PARA: a Data-driven Paratransit Dynamics Model for Planning of the General Santos City Public Utility Jeepney Transport Network

Jasper Francis B. REFUERZO[a], Joaquin Mateo R. JISON[b], Diego Selo L. GLEAN[c], Nestor Michael C. TIGLAO[d], Noriel Christopher C. TIGLAO[e]
[a, b, c, d] *Electrical and Electronics Engineering Institute, University of the Philippines, Diliman, Quezon City 1101, Philippines* [a] *E-mail: jfbrefuerzo@gmail.com*
[b] *E-mail: joaquin.jison@gmail.com*
[c] *E-mail: jameslglean@gmail.com*
[d] *E-mail: nestor@eee.upd.edu.ph*
[e] *National College of Public Administration and Governance, University of the Philippines, Diliman, Quezon City 1101, Philippines*
E-mail: nctiglao@up.edu.ph

Abstract: PARA: a Planning Application for Ridership and Accessibility, is a system that integrates vehicle-agnostic physical sensor data feeds with an analytical machine learning driven model for the per-route operations of a fleet of public paratransit utility vehicles in an urbanized environment. This system integrates with a major ITS pilot test, SafeTravelPH, in General Santos City to collect data and feed it to an open dashboard which can quickly and intuitively provide visualizations and querying tools in order to assist local government units and transport stakeholders with public transport policymaking. The initial test of the system has garnered approval and support from local government units and public utility jeepney cooperatives to assist in their Public Utility Vehicle Modernization Program.
Keywords: Paratransit Modeling, Automatic Vehicle Location Models, Support Vector Machines, Long Short-term Memory, Unsupervised Clustering

1 INTRODUCTION

As is the case in many developing economies around the world, the traffic situation in the Philippines has become a growing concern for policymakers, transportation operators, and passengers alike. The effects are apparent citing estimates by the Japan International Cooperation Agency where economic losses due to congestion alone are as much as 3.5 billion pesos daily in the capital, which may rise to 5.4 billion by 2035. (Staff, n.d.). In 2016, the traffic situation was even declared a national emergency, prompting government-sponsored initiatives to remedy the issue and ensure future improvement.

The problem stems from the under-performance of formal Philippine transportation services, giving rise to ubiquitous, yet unpredictable paratransit operations. Paratransit refers to informal (i.e. privately operated, loosely regulated if not illegal) transportation services which cater to populations that either lack access to public transportation altogether or lack access to sufficient, high-quality service (Shimazaki & Rahman, 1995).

These forms of transportation are also exceedingly common in developing countries around the world, where the development of formal systems has not kept pace with demand for travel (Puspitasari & Maryunani, 2019). When paratransit eclipses formal services entirely, it can create problems for urban planning and the operation of major roadways (Regidor, Vergel, & Napalang, 2009).

While the gradual elimination of paratransit has been suggested as the way forward for developing economies, paratransit operations may also support more formal public transportation through the application of clear regulations and operating zones (Plano, Behrens, & Zuidgeest, 2020). General Santos City is one such area that has begun the process of route rationalization.

General Santos City was primarily chosen for this project due to the city's active and ongoing efforts towards the improvement of its local public transportation. This meant that, at the time of the study, both local government administrators and onground stakeholders such as Public Utility Jeepney (PUJ) operators were available for consultation. This also meant that existing data collection systems had already been deployed in the city for several months, giving the project sufficient historical data. Furthermore, General Santos City, given its status as a first class highly urbanized city, was considered an acceptable model for large, urbanized cities in the Philippines (Elemia, 2016).

In order to assist efforts to define PUJ stops and terminals, an analysis of passenger traffic along the routes is necessary. Such a platform already exists in the form of SafeTravelPH, which operates fleet tracking services for PUJ cooperatives within General Santos at time of writing (N. C. C. Tiglao, Ng, & Tacderas, 2021). This study therefore proposed to extract route-level data such as headway, passenger volume, and origin-destination (OD) matrices from the SafeTravelPH platform to generate a set of parameters for each route which adequately describe the operation of PUJs on such a route.

In line with the objective of promoting data-driven planning, the study provides means of visualizing this information for intuitive interpretation. It also provides means for quickly preparing and executing new scenarios or test cases.

2 RELATED WORK

2.1 Analysis and Planning of Public Transportation

A study by Wu, Ramesh, and Howlett (2015) defines the concept of policy capacity as the "set of skills and resources – competences and capabilities – necessary to perform policy functions". With respect to public transportation in developing countries, research by Perez, Ng, and Tiglao (2022) has used the concept of co-design to retrieve insights directly from stakeholders regarding the most important aspects of policy capacity. These included (among other aspects) the analytical-level capacity to track on-ground traffic conditions, scheduling, stop connections, OD pairs, and travel time estimation, and the operational-level capacity to track driver compliance with routes and legislation.

Their methodology highlighted the use of stakeholder consultations to address the fact that different urban areas may have different priorities with respect to public transportation services. It is evident from these analyses that local government units in the Philippines would benefit from the development of modular tools and services which can help address these identified deficiencies in capacity, while also adapting to local conditions.

A study by Stewart et al. (2016) identifies and visualizes key metrics for transportation systems which are of most significance to planners and regulatory bodies. At the route level, this includes aggregate seating capacity and average bus load. This paper acquired its data from an existing

installation in Montreal, which used Automatic Vehicle Location (AVL) tracking and Automatic Passenger Counting (APC) at bus stops throughout the city. Its presentation of metrics, including density of both vehicles and passengers, strongly informs the outputs desired of the querying system attached to the model this paper aims to develop.

Route rationalization efforts for PUJs were conducted in Baguio City within a similar time frame. A study by Ran˜osa, Fillone, and de Guzman (2017) discusses a methodology employed to characterize PUJ operations in Baguio City. Through methods including volume counts per route, boarding and alighting tracking, passenger origin and destination surveys, the study calculated metrics such as: level of service of the roadway, vehicle load factor, average vehicle speed, OD matrix, and catchment area or area covered by existing transport services.

2.2 Big Data in Intelligent Transport Systems

According to Emmanuel and Stanier (2016), "the term 'Big Data' is in general use to describe the collection, processing, analysis and visualization associated with very large data sets". With respect to ITS, "Big Data" is often used to refer to the data automatically collected through distributed sensor arrays, vehicle tracking, and other such technologies (Kaffash, Nguyen, & Zhu, 2021).

This study will be using data collected from an existing deployment managed by SafeTravelPH; as such, only AVL data including passenger count, boarding and alighting, and vehicle location will be available, with an optimal frequency of one data point per second.

Zhu et al. (2018) classified the applications of Big Data to transport planning into three layers: collection, analytics, and application. Ng, Perez and Tiglao (n.d.) created a parallel framework, in which a cycle of development involving Design Thinking, Analytics, Information Exchange feeds into a larger cycle of improvement in transportation governance. This latter study also highlighted the role of academia in the development of platforms and services without biasing the end product towards desired consumer markets.

Our proposed system acts as the analytics layer in both frameworks, taking the role of creating meaningful insights out of large amounts of data, as guided by stakeholder inputs on which types of insights are most relevant.

A 2019 study by Aranas and Regidor (n.d.) designed a system for automatically recording vehicle speed, location and passenger count on UV Express ("Utility Vehicle Express", a form of minibus) vehicles. The data recovered by such a system would be theoretically compatible with the model this study proposes to develop, and as such the study provides further impetus for the development of such a model.

2.3 Modeling and Simulation of Transport Systems

A key part of this paper is being able to model complex dynamics of a public transport system with the limited data the existing sensors provide. Although there are ways to collect such data by using and implementing more complex and sophisticated sensors, these may not be viable in the current situation.

GPS Tracking and Data acquisition systems are universally implemented in today's smart transit network as part of their AVL systems (Figueiredo, Jesus, Machado, Ferreira, & De

Carvalho, 2001). Their impetus may vary from simple profit-improvement measures to generic information gathering systems. Recent advances have shown their viability in use for transportation planning in modeling traffic signaling, road connection optimization, and route selection in developed countries such as Poland and China (Abduljabbar, Dia, Liyanage, & Bagloee, 2019). Additionally, these systems have been implemented with various vehicle types such as taxis, trucks, and buses. However, these systems are functionally and operationally different from PUJs in the Philippines due their paratransit nature.

Previous research papers have shown the capability of models built upon data gathered from GPS tracking data. Sharman (2014) proposed a method of determining transport system dynamics using GPS data, focusing on the inter-arrival times and activity-state prediction of freight transport. Wei-Hua Lin and Jian Zeng (1999) modeled the inter-arrival times of buses in Blacksburg, Virginia.

Tingting, Gang, Jian, Shanglu, and Bin (2013) were able to create an inter-arrival time system that can take into account multi-route bus systems and bus stop overlap, when normally these would be seen as separate systems in previous models despite being experientially the same for most riders.

Other researchers were able to prove the efficacy of using more modern machine learning techniques. One such paper by Yang, Chen, Wang, Yan, and Zhou (2016) used support vector machines to implement inter-arrival time prediction using road segments with remarkable accuracy with parsimonious training data. Their system, however, is simplistic, only capable of predicting a single pre-learned snapshot of road conditions. This is remedied by a proposal by Qingwen, Ke Liu, Lingqiu Zeng, Guangyan, Lei, and Fengxi (2019) that instead used Recurrent Neural Networks (RNNs), more specifically Long Short-Term Memory (LSTM) network architectures, in predicting inter-arrival times. In doing so they were able to have a system that can take both historical data and newer observed data in order to create more accurate future predictions.

3 PROBLEM STATEMENT AND OBJECTIVES

3.1 Problem Statement

There is a lack of adaptable, accessible tools for use in planning paratransit operations undergoing formalization through the ongoing Public Utility Vehicle Modernization Program in the Philippines. Existing models are largely based on more formal modes of transportation, and are limited in the number of factors which they can account for.
This leads to a lack of policy capacity for public transportation governance.

3.2 Objectives of the Paper

The main objective of this project was to provide a system and model capable of learning and estimating the behaviors of paratransit from a wide array of factors made available from AVL installations already in place. The model outputs should also be accessible in a clear, intelligible format in order to ease its use for planning.

With respect to this objective, the following specifications were set based on key capacity aspects, system parameters, and interface standards discussed in the previous section:

1. The system must be capable of retrieving, reading, querying, and parsing new data as it is recorded directly from the existing SafeTravelPH server platform via a concurrently developed pipeline API.

2. The system must be capable of identifying key locations of passenger flow without any prior contextual knowledge on urban center traffic flows.

3. The system must be able to generate a graph representation of the public transportation network using GPS AVL data.

4. The system must be able to define traffic and flow dynamics of the public transportation system and predict future behavior.

3.3 Scope and Limitations

This paper focuses on the modeling of paratransit traffic behavior based on PUJs but theoretically extensible to similar vehicles. Only the quantities relevant to the calculation of Quality of Service indicators to be defined in the methodology are expected as output.

The interface created for the use of the finished model is not optimized for a highquality user experience Basic usability is the only concern of this aspect of the paper.

This paper only uses data from an existing system (SafeTravelPH AVL feeds). The dataset from this system covers the operations of vehicles from the two largest PUJ cooperatives in General Santos City from January to July of 2021. Note that models of these types appreciate ever-larger amounts of data to improve both accuracy and extensibility.

Finally, this paper focuses on the feasibility of such a system and creating a prototype, as such metrics such as processing speed and optimizations are not a focus so long as they do not significantly impact the user experience.

4 METHODOLOGY

The methodology, seen in Figure 1, is composed of 6 segments. The first 3 segments – data preprocessing, modeling paratransit operations, and modeling system dynamics – are done sequentially. Concurrent with the first 3 segments would be data pipeline design and development of the querying system. Once all 5 segments were completed, they were assembled into a single system. This final system was then presented to stakeholders, consisting of members of the SafeTravelPH system, planning officials of General Santos City, and PUJ transport cooperatives.

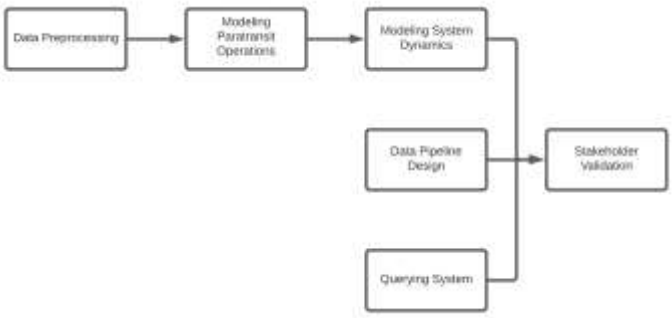


Figure 1: Methodology Flowchart

4.1 Data Preprocessing

SafeTravelPH system data is encoded in a uniform format for entries. This data was not free of errors or invalid values. This is particularly evident in certain trackers having duplicate or no data entries at certain time periods. The following steps were taken in order to allow for data uniformity:

1. Introducing a Uniform UTC Datetime Parameter
2. Dropping Unused Parameters
3. Standardizing Parameter Types
4. Dropping Not a Number (NaN) or Incoherent values
5. Aggregating Timestamps per Second
6. Interpolating missing values between Timestamps

4.2 Paratransit Operations Modeling and Verification

Paratransit Operations Modeling aims to find the following information on how the transit system operates using data-driven analysis techniques:

1. Locations of High Passenger Flow (Stops)
2. Transport Vehicle Flow Directions
3. Identification of Neighboring Stops
4. Route Detection
5. Passenger Flow Statistics per Stop

Locations of high passenger flow will be acquired using a data-driven unsupervised clustering algorithm. Prior studies in passenger volume analysis have shown that density-based clustering algorithms were preferred in passenger volume analysis (Xu & Tian, 2015), as such the following algorithms were chosen for investigation:

1. Density-based Spatial Clustering of Applications with Noise (DBSCAN) (Ester, Kriegel, Sander, & Xu, 1996)
2. Hierarchical Density-based Spatial Clustering of Applications with Noise (HDB-SCAN) (Malzer & Baum, 2020)
3. Ensemble-based Clustering

The resulting algorithm was evaluated by comparing results to the actual locations of key passenger activity determined by the General Santos City local government (ground truth), and by comparing Density Based Clustering Validation (DBCV) scores. For the former, preliminary work on similar data provided by SafeTravelPH showed that a distance MSE of 0.0862 was achievable.

Transport vehicle flow directions, identification of neighboring stops, and route detection will be determined by traversal of data entries per driver per day in order to generate a graph in the form of a stops adjacency list. The per driver per day directed graphs will then be summarized to generate a full graph of the General Santos City PUJ transport network for determining the directions of travel and neighborhoods of stops. Route detection will be conducted based on a K-modes analysis of the adjacency lists per driver, the centroids of which will be classified as detected routes.

Before determining passenger flow statistics, a distinction between pass-throughs and stops for which no passenger activity occurred or was recorded needs to be defined. It is possible that a vehicle did indeed stop at a location, but either no passengers got on or off, or the system failed to record such activity. A support vector machine (SVM) will be used to classify possible stop runs as pass-throughs or as proper stop runs. The following values will be calculated from each run to use as SVM input: total time between first and last point timestamps, total distance covered, maximum acceleration and minimum acceleration. SVM output will be the classification of a run as pass-through or stop. The SVM will be trained on positive stop runs, as well as runs sampled from inter-stop travel points, in order to create clear set identities.

Once this is completed passenger flow statistics per stop will be calculated using simple statistical methods per day per stop.

4.3 System Dynamics Modeling and Verification

System Dynamics Modeling aims to create a graph representation of the transport network with the following information metrics associated per stop:

1. Inter-arrival and Travel Times between any two points on the system.
2. Probability model for a Passenger Boarding and Departing per Stop.
3. Expected Passenger Volume.
4. Predicted Dwell Times per Stop.
5. Travel Time Variation Index (TTVI)
6. Buffering Time Index (BTI)
7. Average Travel Time Index (ATTI)
8. Average to Peak Hour Travel Time Index (APTI)

The first four of these metrics were chosen according to the work of Tiglao, de Veyra and Tolentino as parts of their Service Adequacy factor (N. C. Tiglao, Veyra, & Tolentino, 2019). The

last four of these metrics were based on the Arterial Variation Index defined by Li, Kido and Wang (Li, Kido, & Wang, 2015).

The different dynamics of the system, such as inter-arrival time, is a high-dimensional problem in order to solve but can reasonably be estimated to a decent degree. Many systems already exist in order to model it using linear programming, but are not appropriate in the context of paratransit systems. So these dynamics have to be determined using data-driven techniques. This paper proposes the following models to compare:

1. Sparse Identification of Nonlinear Dynamical systems (SINDy) (de Silva et al., 2020)
2. Support Vector Machines (SVM) (Yin, Zhong, Zhang, He, & Ran, 2017)
3. Long Short-term Memory Recurrent Neural Networks (LSTM-RNN) (Lingqiu et al., 2019)

These models were compared for their accuracy in predicting a test-train split of the data regarding the various model metrics(Botchkarev, 2019). Mean Squared Error (MSE) and Mean Absolute Percentage Error (MAPE) were used to compare these models for their accuracy . An ensemble of these models may also be used and compared.

Once the optimal system model is determined, hyperparameter tuning using grid search will be conducted in order to maximize the accuracy of the chosen model. In the case of SINDy the hyperparameters concerned are a lexicon of non-linear functions and a sensitivity parameter used by SINDy's Sequentially-Thresholded Least-Squares Algorithm (STLSA). As for SVM the hyperparameters are *C*, an optimizer penalty for misclassification, and *gamma*, a closeness parameter. For LSTM-RNNs these hyperparemeters vary depending on choice of structure and optimizer, but most commonly these are the number of layers and nodes in the neural network.

4.4 Data Pipelining

Data Pipelining is an essential portion of this system but can be developed in parallel with the other portions of this methodology. However, a portion of the pipeline depends on both paratransit operations and system dynamics models to be completed and verified. This portion can be broken down to three main application programming interfaces (APIs): an API for SafeTravelPH Data Feeds, an API for Preprocessing, and an API for the Models.

4.5 Querying System

The Querying System should be able to acquire and make use of a subset of the transport network's model in order to make meaningful predictions for the benefit of transport stakeholders. The querying system should be able to use the learned transport network model and provide values for established model outputs given user-defined values for system parameters, including but not limited to:

1. Number of and type of vehicles assigned to a route
2. Time frame, by the hour, day, and/or month.

3. Direct, modified passenger volume data

Processed outputs with accompanying visualizations will also be available. As per prior work of Stewart et al., Tiglao et. al, and Uy et al., these visualizations will necessarily include, but not be limited to, the following, all given per location and timeframe:

1. Passenger Volume, presented as a heatmap or Kernel Density Estimate (KDE)
2. Transport Vehicle Density, presented as a heatmap or KDE
3. Boarding and Alightings, with respect to vehicle and location
4. Vehicle loading factor per route
5. Service area (Catchment), as a simple geometrically defined zone around the routes

4.6 Stakeholder Validation

A substantial success indicator of the designed model is the possible reception by stakeholders; the true value of such a system would be dictated by how adoptable experts and policymakers deem it to be. To further quantify the viability of the designed model, User Experience (UX) usability tests, demonstrations, and a focus group discussion (FGD) will be conducted with a panel of experts, possibly including members of the SafeTravelPH team, in order to acquire evaluations on the usability of the model. Due to time constraints, the group opted to holding an FGD over a survey, as an FGD provides a cost-effective and quick empirical research approach for obtaining qualitative insights, and feedback from practitioners (Kontio, Bragge, & Lehtola, 2008).

5 RESULTS AND DISCUSSION

5.1 Optimal Clustering Algorithm

Three clustering algorithms were initially tested for their viability to provide meaningful clusters to the onboarding and departing data of PUJs from General Santos City, these were: DBSCAN, a Recursive Implementation of DBSCAN, and HDBSCAN.

DBSCAN was able to acquire a large number of stops from initial assessment, as seen in Figure 2, but showed biad towards sparse points, an inability to ”break” large, dense areas in the city center into separate groups, and high variability. A recursive implementation (RDBSCAN), where DBSCAN was run repeatedly on the largest groups, was effective at breaking up large spanning clusters, but was unable to rectify the inability of DBSCAN to grab as much information within the first pass as seen in Figure
3.

Figure 2: Single-pass DB-SCAN Result

Figure 3: Recursive DB-SCAN (Depth 3) Result

Figure 4: Single-pass HDBSCAN Result

All cluster maps above span 22.43 km wide and 28.62 km high.

HDBSCAN is an alternative to the previous algorithms with more transparent hyperparameters (minimum samples and selection ϵ) and methods to prevent singular clustering. The HDBSCAN solution initially had fewer clusters than any of the previous implementations of DBSCAN as seen in Figure 4, however it was able to distinguish between major points of interest in General Santos City, such as cluster separately the two largest malls in the city. Among the three clustering solutions, HDBSCAN is the most promising candidate for clustering, which was doubly confirmed by DBCV scoring, where HDBSCAN (DBCV score of -0.265) outperformed RDBSCAN (DBCV score of -0.801).

5.2 Tuning of Optimal Algorithm

The resulting algorithm after extensive optimization was a three-pass HDBSCAN implementation. Each pass of HDBSCAN was conducted on different sections of the dataset. The first pass identified the largest, densest clusters, while the second used the Open Source Routing Machine (OSRM) API to tag points by their streets, then cluster along streets. The final pass broke clusters wider than 500m into smaller pieces. The hyperparameters of each pass are summarized in Table 1. The minimum samples (s) and minimum cluster size (c) values for the third pass were determined using equations 1 and 2.

Table 1: Three-Pass HDBSCAN Properties

Pass	Min. Sample	Min. Cluster
1st	350	25
2nd	750 (tagged) or 2000 (untagged)	20
3rd	Variable	Variable

$$[h]s = \lfloor min(\frac{size}{2}, 100) \rfloor \quad (1)$$

$$c = \lfloor min(log(size,2),10) \rfloor \quad (2)$$

The tuned HDBSCAN was used to generate both cluster labels for the entire set of identified stop points, as well as a set of stop cluster centroids which would serve as nodes in a graph representation of traffic in the city.

Only similarity to ground-truth data regarding key points for passenger activity was used to verify the output of the optimized model. The results of such comparison, were strongly in favor of the three-pass clustering solution, with a kilometer-based MSE of 5.059 units. When points from sparsely represented areas were excluded, the MSE dropped to 1.34, which was higher than the ideal found in preliminary work but acceptable considering the noise in this data.

5.3 Routing System

The routing system was designed to be able to discover, recreate, and distill route information of different PUJ drivers. A true list of routes is difficult to acquire due to the nature of having multiple PUJ cooperatives, and drivers may choose to ply different routes on different days. This system was meant to acquire such information in an unsupervised manner. The routes for each driver-date pairing were subject to a K-mode clustering (where $k = 10$), to generate the centroids for each of the 10 routes; 4 of these routes are visualized in Figure 5.

(a) Route 1
(7.3 x 3.6km)

(b) Route 2
(29.1 x 14.3km)

(c) Route 9
(29.1 x 14.3km)

(d) Route 10
(7.3 x 3.6km)

Figure 5: Discovered Routes

5.4 Support Vector Machine Pass-through Classifier Implementation

This support vector machine (SVM) model was developed to classify non-stop points within walking distance of stop cluster centroids, here defined as 30m, based on the assumption of a reasonable distance from a point of interest within which pick-ups and drop-offs occur. Such points would be classified as “attempted stops” or “passthroughs” based on driver behavior prior

to the point. For each point, two velocity values measured between two previous points, an acceleration value, and the distance to the nearest stop cluster centroid were used as SVM inputs.

5000 known stops and 5000 inter-stop travel points were sampled from the dataset to be used as training positives and negatives. In the sampling of travel points, data was taken only from the last two months in the dataset (June and July) due to the presence of particular feeds in these months with a high degree of consistency in reporting frequency. As stop points were significantly fewer in number, these were sampled from the entire dataset in order to make up the required number of test points.

The minimum level of accuracy for this part of the system was set at 70%, based on the lower end of accuracy levels achieved by similar SVM-based vehicle behavior classification systems surveyed for reference (Wang, Xi, Chong, & Li, 2017; Chen & Chen, 2017; Karri, De Silva, Lai, & Yong, 2021). The lower value was taken on account of issues with data feed consistency present in the training data.

Testing showed that 85.23% accuracy was achieved on the restricted validation by an SVM with $C = 1$ and a radial basis function (RBF) kernel. The same model achieved 75% accuracy on an unrestricted validation set with negatives sampled from the entire year. The difference may be attributed to poorer data quality, different stopping behavior, or the inclusion of different city areas in previously untrained months. It was nevertheless concluded that this model was acceptable for use in the main data pipeline.

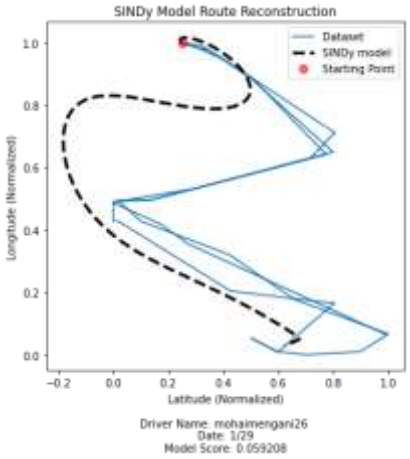
5.5 Sparse Identification of Nonlinear Dynamics

Strict preparation was with the provided data in order for the Python package implementation of SINDy (PySINDy) to function properly. The data set was filtered by driver, date, and month to isolate the routes for modeling. After time conversion, data was aggregated by position and passenger count. All blanks were interpolated for uniform spacing, and all values were normalized.

The remainder of this section presents selected models which highlight the strengths and faults of this approach. For this section, *the score()* function from scikit was used to generate the coefficient of determination (R^2) for each model with respect to the input dataset. All test dates shown in this section are taken from 2021 data.

Figure 6 shows the route of driver *mohaimengani26* as a latitude/longitude plot, with the model reconstruction of this route superimposed. The reconstruction attained a model score R^2 of 0.059208, as the simulated route differed significantly from the measured route. The model of the route of Driver *LADOTRANSCO DRIVER Bantilan* scored 0.622173, and showed much cleaner replication of the original route.

From these simulations, it can be shown that PySINDy had varying effectiveness in modeling routes. Geometrically simpler routes such as that of Driver *LADOTRANSCO DRIVER Bantilan*, whereas more erratic routes such as that of *mohaimengani26* presented less faithful reconstructions.

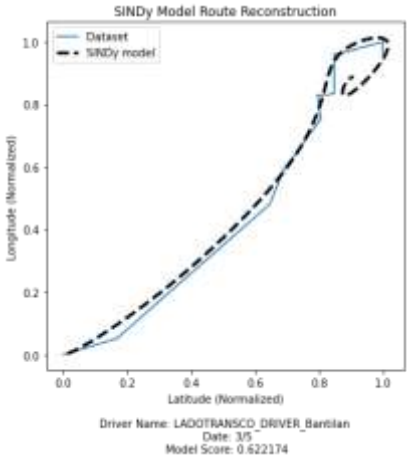


It was also to be noted that PySINDy

$$\dot{lat} = -0.048lat + 0.098long + 0.059lat^2 + 0.004lat * long + -0.066long^2 + 0.044sin(lat) + 0.107cos(lat) + -0.097sin(long) + -0.106cos(long) \quad (3)$$

$$\dot{long} = 0.032lat + -0.035long + -0.027lat^2 + -0.001lat * long + 0.027long^2 + -0.030sin(lat) + -0.045cos(lat) + 0.035sin(long) + 0.045cos(long) \quad (4)$$

Figure 6: mohaimengani26 PySINDy Results



$$\dot{lat} = 0.324lat + -0.201long + -0.427lat^2 + 0.100lat * long + 0.298long^2 + -0.317sin(lat) + -0.699cos(lat) + 0.207sin(long) + 0.699cos(long) \quad (5)$$

$$\dot{long} = 0.196lat + -0.149long + -0.247lat^2 + 0.050lat * long + 0.182long^2 + -0.192sin(lat) + -0.403cos(lat) + 0.155sin(long) + 0.403cos(long) \quad (6)$$

Figure 7: LADOTRANSCO _DRIVER Bantilan PySINDy Results

has difficulty in accommodating loops and overlaps in routes, showing lower R^2 for PUJs that loop more than once along their routes.

Numerous issues arose in both the development and utilization of the PySINDy algorithm. These issues mainly involved the sparsity of the provided measurement data, lapses in data availability, and are summarized in the following points:

1. Frequent generation of near-zero co- 3. Model stability highly sensitive to efficients in models data quality issues which could not all be addressed through processing,

2. Model instability when used with filtering, and initial conditions. *simulate()* function prevented evaluation

The SINDy algorithm showed promising results, but those were overshadowed by the numerous limitations and issues that the algorithm faced. The researchers deemed SINDy unfit for further use in system modelling for this paper.

5.6 Support Vector Regression Model

As they are single-output systems, four SVRs were trained: one each for change in passenger count (pdelta) and time (timedelta) for traveling and dwelling scenarios. Differences in time and passengers were used as correct answers for training samples. Values for timedelta, pdelta, and passenger count were normalized before use. A total of 67707 trips were sampled for training the travel SVRs, while only 39286 points were sampled for training the dwell SVRs.

In parallel, a small neural network used to generate embeddings for the categorical inputs. The results of surrogate model training showed high training accuracy (97.78%), but middling validation accuracy (58.17%). Nevertheless, embeddings generated by this model were found to be more compact and more effective for training than binaryencoded categorical data.

Tests of the most promising SVR candidates for all variables showed validation set MSE no less than 0.65 (in normalized units). The best results for each type are summarized in Table 2.

Table 2: Performance of Best Candidate SVRs, per type

Type	Train MSE	Val. MSE
Travel Time	1.0080	0.6515
Dwell Time	1.1891	0.8598
Travel Passenger	1.1291	1.2259
Dwell Passenger	1.1952	1.1522

Further manual tuning did not reveal a combination of hyperparameters which could improve the testing set MSE beyond 0.65 normalized units, which was achieved with a travel time SVR with $C = 100$ and $\epsilon = 0.01$. SVRs in general are known to deal poorly with count variables, and passenger delta may be modeled as the difference of two count variables: passenger boardings and passenger departures.

5.7 Long Short-term Memory Recurrent Neural Networks

The main issues that determine the topology of the Long Short-term Memory Recurrent Neural Networks (LSTM-RNN) are the significant factor of auxillary categorical inputs, and the nature of the time series being a detection of rare events. Auxillary inputs require a separate model that is attached to the LSTM-RNN, as these are discrete and time independent. Meanwhile for the rare event detection a paper by Uber (Laptev, Smyl, & Shanmugam, 2017) has tested the efficacy of LSTM Autoencoders to generate a general representation of the time series data for rare event detection. Using this knowledge, the encoder half of a Composite LSTM Autoencoder, as seen in

8, is concatenated with a Categorical Embeddings layer as inputs to a feed forward neural network as seen in Figure 9.

The loss function took advantage of huber loss which operates similarly to MSE for values below delta, a hyperparameter for the loss function, and similarly to mean absolute error (MAE) for values above it. The model with embeddings were trained to a huber loss of 0.0754 (equivalent to an MSE of 0.1666) with similarly low results for

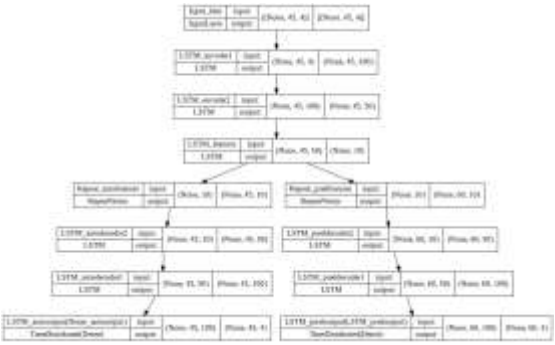


Figure 8: LSTM Composite Autoencoder

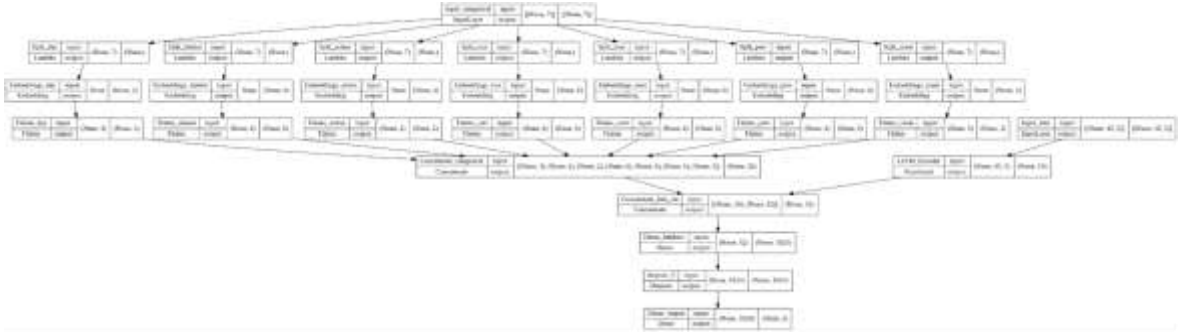


Figure 9: Feedforward Neural Network with LSTM Encoder and Embeddings

the validation set at 0.0780 (equivalent to an MSE of 0.1733). The hybrid model using LSTM Composite Autoencoders and Embeddings was able to garner a high accuracy on the validation set. However, any data feeds with large gaps in time, inconsistent entries, or too few were removed from the dataset used to train this model. The consequences of preparation method are discussed in the following section.

5.8 Model Testing

In order to test the effectiveness of the chosen model with data of more varied quality, additional sample sets were created to stress test each of the models. These sample sets were also created to evaluate the effectiveness of these models for the computation of indices.

Four test sets were prepared with data sampled from the entire year and no intersection with the training sets. These sets included both sparse trips and trips from poorly represented part of the city. The sample sets were as follows:

- Traveling Data, composed of 10109 (10000) trips.

- Dwelling Data, composed of 4999 (5000) trips
- Index Set A, intended for TTVI, BTI and ATTI, composed of 196 points
- Index Set B, intended for APTI, composed of 158 trips

Data was too sparse for data in index sets to be taken from 5-minute time frames as designed in (Li et al., 2015), so six-hour periods were used instead.

Table 3: LSTM Error on Travel and Dwell Testing Sets.

LSTM Error	Travel ΔT	Travel ΔP	Dwell ΔT	Dwell ΔP
MSE	510.22	38.53	617.40	21.13
MAPE	1.96	2556.78	0.97	2755.63

The results for timedelta and passenger delta prediction for both dwelling and travel data testing sets are shown in Table 3. These results were poor in comparison to results from validation sets. This significant degradation in performance was likely due to the LSTM incorrectly encoding data points that required interpolation. Additionally not all possible pairings of routes, stops, and dates were accounted for due to the smaller dataset.

Table 4: SVR Error on Travel and Dwell Testing Sets.

SVR Error	Travel δT	Travel δP	Dwell δT	Dwell δP
MSE	583.33	20.42	1209.68	12.37
MAPE	0.80	258.50	0.95	301.23

In comparison, the performance of the SVR model, as shown in Table 4 did not degrade to the same degree as that of the LSTM. The most likely reason identified for this is the larger, less strictly filtered training set used by the SVR. A greater exposure to the variety and inconsistency present in the dataset overall allowed this model to cope more effectively with the testing set.

Table 5: Error values for LSTM on Index Testing Sets.

LSTM Error	TTVI	BTI	ATTI	APTI
MSE	0.356	0.590	902.726	0.087
MAPE	0.776	0.357	2.121	3.748
Typical Value Ranges	0.2-0.5	0.2-0.8	1.5-3	0.2-0.5

Table 6: Error values for SVR on Index Testing Sets.

SVR Error	TTVI	BTI	ATTI	APTI
MSE	0.460	0.454	455.763	0.413
MAPE	0.770	0.246	0.673	1.417
Typical Value Ranges	0.2-0.5	0.2-0.8	1.5-3	0.2-0.5

The apparent better performance in index test results was only a consequence of indices generally possessing small ranges of possible values. When compared to expected values for these indices, it is clear that model performance in these tests was equally poor. While these results did prove that the models were capable of generating predicted values for these indices, the level of

error present in the predictions indicates that generated index values will be of limited use until model testing accuracy is also improved.

5.9 Query System and Interface Development

Query system development proceeded in three phases: the extraction of necessary summary statistics, the development of an API to serve queries, and the development of an interface to provide interactivity and visualization to the API.

Summary statistics were generated for each stop, including: average dwell time, average boardings per day, and average departures per day. For each connection between stops the following additional data was generated: estimated load factor, average travel time, average boardings per day, and average departures per day. An average maximum PUJ capacity of 20 people was used to calculate the load factor. All summary statistics extracted in this way were stored to a single CSV file stored to the root.

In order to serve analyzed and generated data to an end user, an API was developed in Flask for Python 3. The developed Flask server is capable of providing the following forms of information, organized by type:

1. Density Data: Passenger and vehicle counts with location and time
2. Stop Clusters: Stop cluster centroid locations, average boardings, departures, and dwell times per stop
3. Stop Connections: connections between centroids, average boardings, departures, and travel times
4. Route Estimation: connection to prediction models from previous sections, allowing forecasting of route behavior
5. Discovered Routes: routes discovered through application of k-modes and associated metrics.

Endpoints associated with this data were also capable of accepting filter keywords such as day, time, and driver when applicable.For the purposes of development, this server was hosted on the free Heroku web hosting platform.

Figure 10: Heatmap Module of Interface, showing density of PUJ stops.

An independent interface was developed in Vue 3, and contains four modules corresponding to API endpoints: Heatmap, Stops, Routes, and Discovered Routes. These modules were also developed with the intention of matching specific interface goals set in the methodology.

1. Heatmap: displays Passenger and vehicle density in filtered ranges as a scaled interactive heatmap. This module is presented as a sample in Figure 10.
2. Stops: displays a map of General Santos with interactive stop cluster centroids, catchments, and connections overlaid.
3. Routes: gives direct access to model predictions for hypothetical trips.
4. Discovered Routes: displays discovered routes as graphs over city streets, with associated data.

5.10 Stakeholder Consultation and Evaluation

Stakeholders from General Santos City, including representatives from PUJ cooperatives, city LGU officials, representatives from the local universities, and staff and operators of SafeTravelPH, were all consulted for vital insights on the end-user perspective for this project.

A preliminary meeting was held on April 7, 2022 with members of the SafeTravelPH team and stakeholders in the transport field of General Santos City, including fleet operators, public safety office members, local government officials, and other key figures in the public transport sector of the city. This meeting acted as an introductory discussion for our work, as well as an initial feedback forum where stakeholders could provide insights and suggestions that would be relevant for further development of the project.

The following points summarize the results of the consultation:

1. Identification of key passenger activity points was of enough relevance that the local government had already done a separate study on it
2. Features defined in the initial objective list were in line with the desired use cases of stakeholders.

The first point highlighted the value of a system which could retrieve key passenger activity locations from AVL data, and indicated that this feature would be an asset to a generalized system for cities which had not yet done such a study. The second indicated that this system does possess features which are relevant to a wide variety of stakeholders associated with transportation policy. Overall, this consultation confirmed the relevance of this system as a tool for expanding policy capacity in the transport sector, especially given that it is flexible enough to be applied to data from other cities.

On June 14, 2022, a usability test for the final system prototype was held with select members coming from the SafeTravelPH Team, General Santos City LGU, Local PUJ Cooperatives, and the Local Public Safety Office, totalling 8 participants. This meeting served as a culminating demonstration and discussion regarding the PARA online tool interface.

The program started with an introduction of the tool as well as the background and motivations behind the development of the tool. Next, a walkthrough of all the interface features was presented. The researchers went through every feature present in each prototype interface, and demonstrated the functionality and possible uses of each feature. Guide questions and sample scenarios (e.g., ”Provided a route by the presenters, which linkages along this route are the slowest?”) were

laid out during this walkthrough in order to encourage more audience participation during the program.

Once the demonstration had concluded, and the participants had experience exploring the interface of the tool, a focus group discussion and Q&A portion was conducted in order to acquire the participants' feedback. From this discussion, the following insights and recommendations were gathered:

- Information about reconstructed routes was found greatly useful for monitoring driver compliance with approved routes.
- Future modules should include insights on fuel consumption along routes.
- Plotting prescribed LPTRP routes against data acquired routes would be useful.
- An added feature where the node changes color based on passenger activity intensity within the current time range would be helpful.
- Maximum, minimum and percentile values would be more useful than average dwell time as an absolute value.
- Adopting these modules for tricycle operation could be helpful.
- Plotting formal boarding and departing points would be helpful.
- There exists an ordinance which limits PUJ layovers to 3 minutes; monitoring driver compliance with this ordinance would be helpful.

Public Transport Alliance of GenSan (PTAG) members and PUJ operators were overall welcoming in their response to these tools, and representatives at the usability test expressed hopes that the LGU and PSO may utilize these tools in traffic enforcement deployments.

After the usability test, a small survey was conducted to measure the participants' reception of the developed tool. This survey included the following statements, with the responses being numbers on a scale from 1 to 5, corresponding to "strongly disagree" to "strongly agree", respectively:

1. The interface has features which are useful to me.
2. The interface is easy to use.
3. The charts and graphs in this interface present meaningful data.
4. The descriptive information about points and trips is meaningful.
5. The route prediction feature provides meaningful insights.
6. I would use a completed version of this interface for my work.

This survey was answered by the participants of this meeting, totalling eight responses. The six survey questions correspond to six metrics: features, ease of use, visual relevance, data relevance, route prediction value, and desire to use. Figure 11 shows the average ratings per metric.

The stakeholders' evaluations rated the list of features and the desire to use the completed tool highly, with averages of 4.375 out of 5. The visual relevance and route prediction metrics follow with ratings of 4.25 out of 5. ease of use and data relevance scored the lowest, with an average of 4 out of 5.

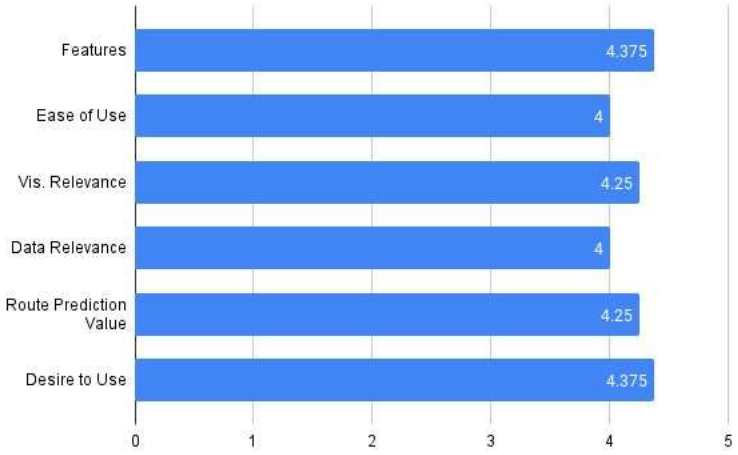


Figure 11: Usability Test Survey Results

Improvements could still be made in regards to the presentation and the relevance of the data, but despite that, the participants highly regarded the features of the tool, and were looking forward to using it in the near future. Ease of use was generally lacking, but that is to be expected since the scope of the project only covers a simple interface.

Lastly, the audience took part in an activity which served as the closing segment. Using the interface, the participants were tasked to find certain metrics and values given certain scenarios.

This final usability test confirmed that the system developed has a number of direct applications with respect to improving policy capacity in General Santos City and potentially in other urbanized cities in the country. In particular:

1. Systems for characterizing travel and dwell time along existing routes have direct potential for assisting **city planning** with the assessment of these routes.

2. Systems for characterizing passenger stop demand density have potential for assisting **city planning** with the placement of waiting or mode-switching infrastructure.

3. Systems for route reconstruction and driver behavior identification have uses with **PUJ fleet operators** for monitoring driver compliance with local policy.

4. Systems for estimating trip time may be useful for **passenger** information systems.

Additionally, the structure of the system outputs (an API with a separate front-end) allows interfaces and the modules present on them to be customized to the needs of a particular city or stakeholder; additionally, expansions to the kinds of data provided are readily accomplished by adding more routes to the API.

Maintained deployment to General Santos City was not within the scope of this project, however development and maintenance of the system is expected to continue through associated research institutions such as the Ubiquitous Computing Laboratory of the University of the Philippines. Extensive documentation and modular software design (to be discussed in the next section) were intended to ease the process of further development.

5.11 Data Pipeline

Initially the data pipeline was designed to be a rigid system that accepts raw data that is in real time. However, after consultation with stakeholders and partners at SafeTravelPH, this design was scrapped in favor of a flexible modular API system. Primarily because the pipelined systems, such as the data feeds or the querying system, may have different input and output data feed shapes as data sources or outputs change. Additionally, modularizing the different processes allow easier replacement with updated models or removed entirely if data processing needs change.

In response to this, the following processes were converted into separate standalone python modules and ipython notebooks:

1. Data Input Processing from Github
2. Data Input Processing from local CSV files
3. Data Processing, Standardization, and Aggregation
4. Three-pass HDBSCAN Clustering for Stops
5. Project OSRM Street Tagging
6. Visualization of Cluster Centroids
7. Generator for Transport Network Graph
8. K-modes Route Discovery
9. SVM for Passthrough Classification
10. PySINDy for Prediction
11. SVR for Prediction
12. LSTM-RNN for Prediction
13. Prediction Test Set Generation and Validation

These modules were used to generate, processes, and feed data into the Flask API as described in Section 5.9 for use in the Querying system.

6 CONCLUSION

6.1 Conclusion

The majority of research and development into intelligent transport systems are built upon the backbones of high investment infrastructure as well as planned and regulated public transportation systems. Developing countries cannot share in these developments due to a large share of their public transport network served by informal paratransit.

With the assistance and support of the local government of General Santos city and the team of SafeTravelPH, PARA was able to prove itself as a means to improve policy capacity by formalizing and visualizing vehicle-agnostic data. In this case, the results were successfully

generated from data collected from PUJ transport cooperatives to support their Public Utility Vehicle Modernization Program. Through this, a positive step was made on the road to developing robust intelligent transport systems across the Philippines.

PARA presents itself as a way to bridge the gap and serve as the beginning of future developments in data driven systems to assist in understanding and planning paratransit systems in the Philippines, as well as other similar countries that rely on an informal sector for their public transportation needs.

6.2 Recommendations

The following are recommendations for future work based off of this paper:

1. Use larger, more consistent data feeds that span a longer amount of time.
2. Explore other PySINDy function libraries and optimizers, such as SSR and SR3, in order to improve accuracy.
3. Improve or augment dataset with more features (such as speed and traffic conditions), and use Principal Component Analysis to maximize the accuracy and resolution offered by neural networks.
4. Evaluate possible effects of locality factors such as topography, zoning, and demographics as features that affect passenger load and travel time.
5. Predict other metrics such as fuel consumption, connections to other modes of transportation, and law or policy enforcement conditions.
6. Explore and characterize driver behavior with a longer time history to improve predictions of passthroughs and the LSTM.

References

Abduljabbar, R., Dia, H., Liyanage, S., & Bagloee, S. A. (2019). Applications of artificial intelligence in transport: An overview. *Sustainability*, *11*(1), 189.

Aranas, D. C. D., & Regidor, J. R. F. (n.d.). Camera-aided traffic data collection method for paratransit services in the philippines.

Botchkarev, A. (2019). A new typology design of performance metrics to measure errors in machine learning regression algorithms. *Interdisciplinary Journal of Information, Knowledge, and Management*, *14*, 045–076. Retrieved from http://dx.doi.org/10.28945/4184 doi: 10.28945/4184

Chen, H., & Chen, L. (2017). Support vector machine classification of drunk driving behaviour. *International journal of environmental research and public health*, *14*(1), 108.

de Silva, B. M., Champion, K., Quade, M., Loiseau, J.-C., Kutz, J. N., & Brunton, S. L. (2020). Pysindy: A python package for the sparse identification of nonlinear dynamical systems from data. *Journal of Open Source Software*, *5*(49), 2104.

Retrieved from https://doi.org/10.21105/joss.02104 doi: 10.21105/joss .02104

Elemia, C. (2016, 9). *Senate wraps up hearings on emergency powers vs traffic.* Retrieved from https://www.rappler.com/nation/147027-senate-emergency-powers-passage-december/

Emmanuel, I., & Stanier, C. (2016). Defining big data. In *Proceedings of the international conference on big data and advanced wireless technologies* (pp. 1–6).

Ester, M., Kriegel, H.-P., Sander, J., & Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In (p. 226–231). AAAI Press.

Figueiredo, L., Jesus, I., Machado, J. T., Ferreira, J. R., & De Carvalho, J. M. (2001). Towards the development of intelligent transportation systems. In *Itsc 2001. 2001 ieee intelligent transportation systems. proceedings (cat. no. 01th8585)* (pp. 1206–1211).

Kaffash, S., Nguyen, A. T., & Zhu, J. (2021). Big data algorithms and applications in intelligent transportation system: A review and bibliometric analysis. *International Journal of Production Economics*, *231*, 107868.

Karri, S. L., De Silva, L. C., Lai, D. T. C., & Yong, S. Y. (2021). Identification and classification of driving behaviour at signalized intersections using support vector machine. *International Journal of Automation and Computing*, *18*(3), 480–491.

Kontio, J., Bragge, J., & Lehtola, L. (2008). The focus group method as an empirical tool in software engineering. In F. Shull, J. Singer, & D. I. K. Sjøberg (Eds.), *Guide to advanced empirical software engineering* (pp. 93–116). London: Springer

London. Retrieved from https://doi.org/10.1007/978-1-84800-044-5 _4 doi: 10.1007/978-1-84800-044-5 _4

Laptev, N., Smyl, S., & Shanmugam, S. (2017). Engineering extreme event forecasting at uber with recurrent neural networks. *UBER Engineering*, *9*, 06–17.

Li, R., Kido, A., & Wang, S. (2015). Evaluation index development for intelligent transportation system in smart community based on big data. *Advances in Mechanical Engineering*, *7*(2), 541651.

Lin, W.-H., & Zeng, J. (1999). Experimental study of real-time bus arrival time prediction with gps data. *Transportation Research Record*, *1666*(1), 101-109.

Retrieved from https://doi.org/10.3141/1666-12 doi: 10.3141/1666-12

Lingqiu, Z., Guangyan, H., Qingwen, H., Lei, Y., Fengxi, L., & Lidong, C. (2019). A lstm based bus arrival time prediction method. In *2019 ieee smartworld, ubiquitous intelligence & computing, advanced & trusted computing, scalable computing & communications, cloud & big data computing, internet of people and smart city innovation (smartworld/scalcom/uic/atc/cbdcom/iop/sci)* (pp. 544–549).

Malzer, C., & Baum, M. (2020, 9). A hybrid approach to hierarchical density-based cluster selection. *2020 IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems (MFI)*. Retrieved from http://dx.doi.org/10.1109/MFI49285.2020.9235263 doi: 10.1109/mfi49285.2020.9235263

Ng, A. C., Perez, R., & Tiglao, N. C. C. (n.d.). Building policy capacity of local governments for big data applications in public transportation.

Perez, R. E., Ng, A. C. L., & Tiglao, N. C. C. (2022). Enhancing policy capacity through co-design: the case of local public transportation in the philippines. *Policy Design and Practice*, *5*(1), 103–121.

Plano, C., Behrens, R., & Zuidgeest, M. (2020). Towards evening paratransit services to complement scheduled public transport in cape town: A driver attitudinal survey of alternative policy interventions. *Transportation Research Part A: Policy and Practice*, *132*, 273–289.

Puspitasari, E., & Maryunani, P. (2019). The analysis of problem solving priority of semi paratransit services in developing countries. In *11th asia pacific transportation and the environment conference (apte 2018)* (pp. 32–40).

Ran˜osa, L. L., Fillone, A. M., & Guzman, M. P. D. (2017). Jeepney service operation and demand in baguio city, philippines..

Regidor, J. R. F., Vergel, K. N., & Napalang, M. S. G. (2009). Environment friendly paratransit: re-engineering the jeepney. In *Proceedings of the eastern asia society for transportation studies vol. 7 (the 8th international conference of eastern asia society for transportation studies, 2009)* (pp. 272–272).

Sharman, B. W. (2014). Behavioural modelling of urban freight transportation: Activity and inter-arrival duration models estimated using gps data..

Shimazaki, T., & Rahman, M. M. (1995). Operational characteristics of paratransit in developing countries of asia. *Transportation research record, 1503*, 49.

Staff, C. P. (n.d.). *Ph traffic may worsen, to cost ‘ 5.4 billion daily - jica.* Retrieved from https://cnnphilippines.com/news/2018/09/19/JICA-study-traffic-5-billion.html

Stewart, C., Diab, E., Bertini, R., & El-Geneidy, A. (2016). Perspectives on transit: Potential benefits of visualizing transit data. *Transportation Research Record*, *2544*(1), 90-101. Retrieved from https://doi.org/10.3141/2544-11 doi: 10 .3141/2544-11

Tiglao, N. C., Veyra, J. M. D., & Tolentino, N. J. Y. (2019). The quality of service perception among public transport users in metro manila considering dominance of paratransit modes..

Tiglao, N. C. C., Ng, A. C. L., & Tacderas, M. A. Y. (2021). *Supporting collaborative governance and city-wide public transport reforms in general santos city, philippines through crowdsourcing and digital co-production.* (Unpublished Manuscript)

Wang, W., Xi, J., Chong, A., & Li, L. (2017). Driving style classification using a semisupervised support vector machine. *IEEE Transactions on Human-Machine Systems*, *47*(5), 650–660.

Wu, X., Ramesh, M., & Howlett, M. (2015). Policy capacity: A conceptual framework for understanding policy competences and capabilities. *Policy and Society*, *34*(34), 165–171.

Xu, D., & Tian, Y. (2015). A comprehensive survey of clustering algorithms. *Annals of Data Science*, *2*, 165—193.

Yang, M., Chen, C., Wang, L., Yan, X., & Zhou, L. (2016). Bus arrival time prediction using support vector machine with genetic algorithm. *Neural Network World*, *26*(3), 205.

Yin, T., Zhong, G., Zhang, J., He, S., & Ran, B. (2017). A prediction model of bus arrival time at stops with multi-routes. *Transportation Research Procedia*, *25*, 4623-4636. Retrieved from https://www.sciencedirect.com/science/ article/pii/S2352146517306889 (World

Conference on Transport Research WCTR 2016 Shanghai. 10-15 July 2016) doi: https://doi.org/10.1016/j.trpro .2017.05.381

Zhang, C., & Teng, J. (2013). Bus dwell time estimation and prediction: A study case in shanghai-china. *Procedia-Social and Behavioral Sciences*, *96*, 1329–1340.

Zhu, L., Yu, F. R., Wang, Y., Ning, B., & Tang, T. (2018). Big data analytics in intelligent transportation systems: A survey. *IEEE Transactions on Intelligent Transportation Systems*, *20*(1), 383–398.