Characterization of the Spatiotemporal Behavior of a Sweeping System Using Supervised Machine Learning Enhanced with Feature Engineering

Bechir Ben Daya, Jean-François Audy, Amina Lamghari

▶ **To cite this version:**

Bechir Ben Daya, Jean-François Audy, Amina Lamghari. Characterization of the Spatiotemporal Behavior of a Sweeping System Using Supervised Machine Learning Enhanced with Feature Engineering. 23th Working Conference on Virtual Enterprises (PRO-VE), Sep 2022, Lisbon, Portugal. pp.245-261, 10.1007/978-3-031-14844-6_20 . hal-04642024

HAL Id: hal-04642024

https://inria.hal.science/hal-04642024v1

Submitted on 9 Jul 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons Attribution 4.0 International License

This document is the original author manuscript of a paper submitted to an IFIP conference proceedings or other IFIP publication by Springer Nature. As such, there may be some differences in the official published version of the paper. Such differences, if any, are usually due to reformatting during preparation for publication or minor corrections made by the author(s) during final proofreading of the publication manuscript.

Characterization of the Spatiotemporal Behavior of a Sweeping System Using Supervised Machine Learning Enhanced with Feature Engineering

Bechir Ben Daya, Jean-François Audy and Amina Lamghari,

Business School, UQTR (3351, boulevard des Forges, Trois-Rivières (Québec) G8Z 4M3, Canada)

bechir.ben.daya@uqtr.ca, jean-francois.audy@uqtr.ca, amina.lamghari@uqtr.ca

Abstract. This paper focuses on geolocation data processing to infer the behavior of a mechanical sweeping system. A framework based on the feature engineering (FE) and machine-learning (ML) tools for geolocation data processing is proposed. A supervised multi-classification machine learning using a large range of classifiers, input variables, training and data test sets is used to predict the sweeping system behavior. The results showed that Logistic Regression (LR) and Support Vector Machine (SVM) are the best classifiers for predicting the sweeping behavior and some simulated instances constituted the best training sets. The sweeping state prediction accuracy provided with LR and SVM classifiers, when trained with historical data, were in average 86.22% and 86.13%, respectively. These predictions using the same classifiers, when trained with simulated data, were in average 87.40% and 87.22%. These promising results illustrate the potential of integrating FE and simulation to enhance the performance the ML tools when studying the behavior of complex logistics systems.

Keywords: Supervised machine learning, Feature engineering, Multi-classification, Big Data processing, Geolocation data, Sweeping system.

1 Introduction

Every spring in Canada, a considerable amount of abrasive material applied during winter road maintenance is removed from the road network by mechanical sweeping to increase road safety and reduce environmental impacts. Recently, a small-sized enterprise designed and manufactured a novel broom which significantly changes the road sweeping logistics. Up to known, no evaluation has benne carried out in terms of operational and environmental performance for this novel broom mode. To evaluate and ultimately improve the sustainability of the sweeping system as well as for its virtualization based on AI applications for smart city (e.g., provide information about the progression of the city services to the citizens), a large amount of geolocation data was collected. The streaming data creates its own challenges in terms of how to process the large volume of information collected to determine the simulation parameters while ensuring quality and accuracy. Indeed, rather than visualizing over 400 hours of recorded videos, a data analysis approach was developed. The methodology adopted

consists in an approach for processing GPS data using Feature Engineering (FE) and machine learning (ML) tools to identify the behavior of the system in order to compute the input parameters for a simulation model. A sample of data was first processing and validating manually for later use as the training and test data sets. In a second step, this data will be used to train classifiers in order to predict, using ML tools, the sweeping behavior over all the collected data.

To achieve the characterization of the sweeping states and its attributes, a wide range of classifiers, input variables and validated training data sets are used. The results show that Logistic Regression (LR) and Support Vector Machine (SVM) are the best classifiers and some simulated instances constituted the best training sets. The sweeping state prediction accuracy provided with LR and SVM classifiers when trained with simulated data were in average 87.4% and 87.22%.

This paper has the following contributions to the study of behavior geolocation data processing using FE tools. First, we propose a smoothing heuristic to the raw data preprocessing stage in order to clearly identify and separate the sweeping states. Second, we propose a classification framework based on two steps, rather than one. The first step classifies the system's states based on the speed variable using appropriate thresholds while the second step makes use of several input variables provided by the first step in its classification. This leads to substantial improvements in the performance of the classification scheme. Third, the accuracy of the prediction using a simulated training data instances gives a better and more stable results than when using historical data. Fourth, a corrective heuristic was proposed to improve the classification of the states of short duration leading to much more accurate results.

The remainder of this paper is organized as follows. A literature review is presented in Section 2. Section 3 outlines the methodology followed while Section 4 deals with the model building and its application. Finally, a conclusion is provided in Section 5.

2 Literature Review

The deployment of the concept of connected vehicles provides a large volume of geolocation data tracked with GPS technology, as noted in [1]. In [2], authors suggested that GPS data analysis could provide a better characterization of the spatiotemporal movement of vehicles. However, the scale of ingested data in the transportation system has become a bottleneck for the traditional data analytics solutions as reported by [3]. ML tools provide data-driven solutions that can cope with the new analysis requirements. It is also noted that the application of ML tools can be used to identify the purpose of a trip and mode of travel on GPS trajectory data according to [4]. However, despite the fact that GPS trackers provide valuable data, they fall short in terms of describing the behavior of a complex system faithfully as noted in [5, 6]. Diverse opportunities to enhance data analytics and applications for logistics and supply chain management, including technology-driven tracking strategies are considered in [7]. Feature engineering tools is an essential discipline to improve the performance of prediction models applied to GPS data to infer the behavior of logistics

systems. As noted in [8], FE is a crucial step in the predictive modeling process. It includes building new features from the given data in order to enhance predictive learning performance as suggested in [9]. Similarly, in [10] the author noted that FE as "the task of improving predictive modelling performance on a dataset by transforming its feature space". This discipline "involves domain knowledge, intuition, and most importantly, a long process of trial and error" as reported by [8].

Next, we review the literature dealing with the processing of geolocation data approaches to infer the behavior of transport vehicles.

Using GPS tracking and accelerometer data, authors in [11] focused on how to improve the trip purpose identification technique. The results show that Random Forests (RF) provide robust trip purpose classification with correct predictions between 80% and 85%. This work indicated that ML tools could enhance GPS data using classification in the case of a repetitive trip when data sets used are susceptible to learning. However, for non-repetitive contexts, one can only determine the mode or the state of the system. An innovative methodology for inferring process states from geolocation data is proposed in [12]. Geolocation data can be used to get insight into transportation processes, operations, and service quality, as noted by the authors. The methodology proposed uses the zero-speed threshold to identify stationary and non-stationary events from geolocation data. However, this methodology is more effective in the case where the processes are highly structured and the behavior is well defined and predictable. In [13], authors analyzed the vehicle behavior and extracted operational information using the segmentation of GPS trajectory data generated in logistics transportation. The main contribution is the layout design of convolutional neural network input layer, which represents the fundamental motion characteristics of a moving object including speed, acceleration, jerk, and bearing rates. A highest accuracy of 84.8% has been achieved. This methodology has also been contrasted with traditional ML algorithms.

In [14], authors focused on the segmentation of GPS trajectory data generated in logistics transportation in order to extract operational information related to the vehicle behavior characteristics. The authors noted that the widely applied ML technique K-Nearest Neighbors (KNN) is used to tackle the same trajectory data segmentation problem. The precision and recall for KNN are both 86% to recognize that stopping points are business points. Although KNN performs better in precision compared with probabilistic logic data segmentation problems, it cannot filter out all the real business points. In [4], the application of ML methods to identify the purpose of a trip and mode of travel on GPS trajectory data shows that RF method is more efficient than the decision tree method. The RF and decision tree methods have already proven to be better than some of the other supervised ML methods for the identification of trip purpose as noted in [15].

This literature review shows that very few studies focused on predicting the behavior of an object tracked by GPS using ML, while even fewer studies presented GPS data processing approaches using a FE framework. To the best of our knowledge, no work has addressed the behavior of complex logistics system such as a sweeping system using geolocation data and ML tools with a wide range of multi-classification supervised algorithms. Furthermore, no work has presented a framework based on

clustering and multi-classification applied in two successive steps to process real data using classifiers trained with simulated data.

3 Methodology

The study of the sweeping system behavior to evaluate its carbon footprint, to improve its performance, and for its possible virtualization within the framework of a smart city, requires the collection and the processing of real data. To collect the necessary data needed, a set of cameras (front and rear camera) with embedded GPS were installed in the brooms and trucks involved in the sweeping system during a full season of operation. The processing of this large volume of collected data, posed a challenge and call for the development of appropriate methods capable of processing geolocation data in order to infer the behavior of the sweeping system. The methodology adopted consists of developing an approach for processing GPS data using ML tools to identify the behavior of the sweeping system in order to produce the input parameters required for a simulation model and for potential AI-based applications.

3.1 Description of the Sweeping System

The mechanical sweeping system employs a principal broom that continuously loads a dump truck, which is followed by a traditional broom for finishing, as shown in Fig. 1. The behavior of the mechanical sweeping system consists of alternating between different states of sweeping, waiting, or moving:

- *Sweeping state*: when the broom is performing a sweeping operation;
- *Waiting state*: when the broom is in standby mode for various reasons;
- *Moving state*: when the broom is travelling to/from a sweeping area without performing sweeping operations.

Fig. 1. The mechanical sweeping system considered by this study

3.2 Data Processing

A FE framework based on a two-step classification-clustering scheme is conducted to infer the correct behavior of the sweeping system. Figure 2 describes the steps used for data collection and processing. To achieve this objective, we subjected the geolocation data to a set of processing steps. Data collection is followed by data preprocessing, which aims to clean the data by fixing any record errors, especially with regards to time or distance, and splitting this data into different shifts. The objective of the next step is to manually process a batch of data to be used for training and testing the ML classification models. The last two steps relate to the building of an ML classification model and its application to predict the sweeping behavior from the collected data. The classification resulted in a database of sweeping states sorted by their attributes in terms of duration and average speed.

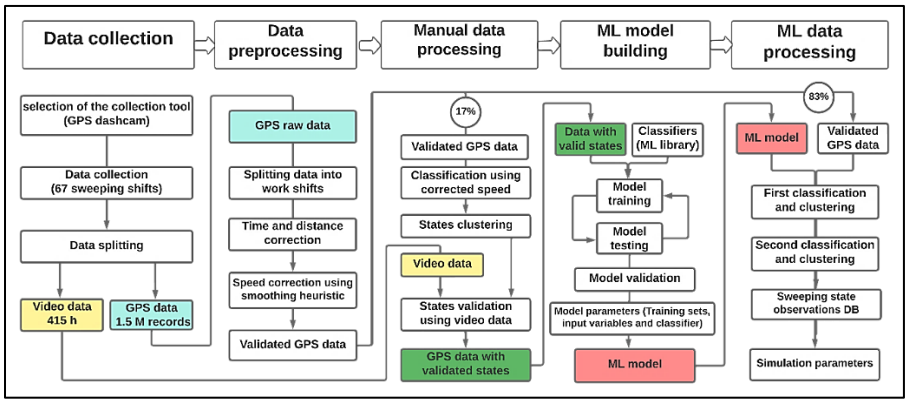


Fig. 2. Geolocation data processing approach reinforced by FE

Data Collection

During the 2019 spring season, over 400 hours of recorded video were collected from 67 work shifts for a total size of 3.5 Tb. From the different videos, we have extracted geolocation data. This task was enabled through Dashcam viewer software Version 3.3.2 that extracts maps into KML files and GPS data into CSV files with structured data. Each second of video recording corresponds to a line of the described data and thus generated nearly 1.5 million lines of GPS data.

Data Preprocessing

Data preprocessing involves error and speed correction.

Error correction

Geolocation data is referenced according to the corresponding work shift and is then structured as a geolocation Excel database. Some errors in recording geolocation data were found. These errors were related to odometer initialization when the broom engine is switched off, some missing recordings of various duration, some redundant

recordings, and overlap between shifts. These errors were fixed using appropriate manual processing schemes.

Speed Correction

The speed curve of one shift illustrated in Fig. 3 shows that, when adopting the speed threshold for classification, a large number of states would be obtained with a short and insignificant duration due to the alternating speed between adjacent states. For example, when the broom is in the sweeping state, its speed may increase and cross the threshold between the sweeping and the moving states for a short duration (Fig. 4.a). In this case, the speed classification will produce an alternation between these two states while the broom is always in the sweeping state. The most frequent overlaps are illustrated in fig. 4 a-d. To remedy this possible misclassification, states that have a small duration should not be classified as separate states, but rather as part of adjacent states. A smoothing heuristic, which will be presented later, is applied to correct the speed in order to minimize these overlaps.

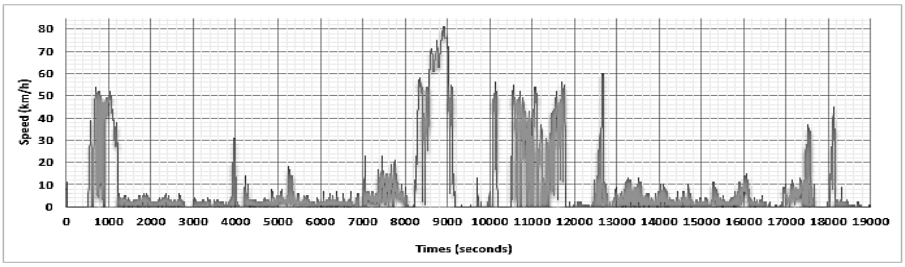
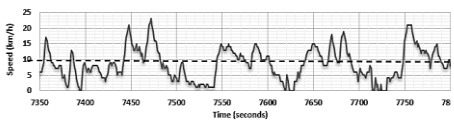
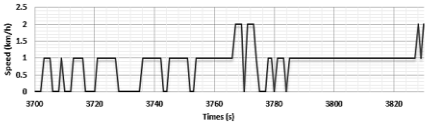


Fig. 3. Speed curve of one shift having a duration of 19000 s

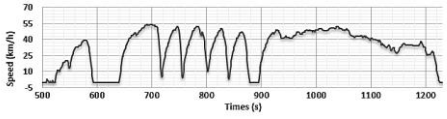
Broom speed differs depending on its state. In the waiting state, the speed is zero, as noted in [14], while in the sweeping state, the speed is lower than in the moving state. In fact, the speed threshold can be used to characterize the various broom states, e.g., in [16], authors classified interstate data into the driver behavior (slow, normal, aggressive) based on the speed thresholds.



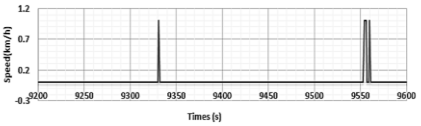
(a) Sweeping high speed



(b) Sweeping low speed



(c) Moving acceleration-deceleration and stop in traffic



(d) Noise (GPS error) in the waiting state

Fig. 4. The remarkable overlaps in the speed curve

In order to partially eliminate the interference between states, a smoothing heuristic is built to adjust the speed of each record. This is done by exploring forty-second-records (ten backward and 30 forward around each record). The 40 seconds duration of the exploration zone is based on experimentation and estimates to the 2/3 of the minimum of the state duration considered. The average speed of each zone is used to adjust the speed of that record in order to minimize the possible overlaps. This heuristic is illustrated in Fig. 5 where S_i is the speed of the record under consideration and AS_i is the average speed of the 40 records around the current record. The adjustment carried out by the heuristic is based on the thresholds between the three states, as mentioned earlier. However, the threshold between waiting and sweeping is adjusted from 1 to 0.4 based on extensive experimentation in order to minimize the number of states in the characterization phase.

ℓ = threshold between waiting and sweeping
L= threshold between sweeping and moving
CS_i: corrected speed for record i

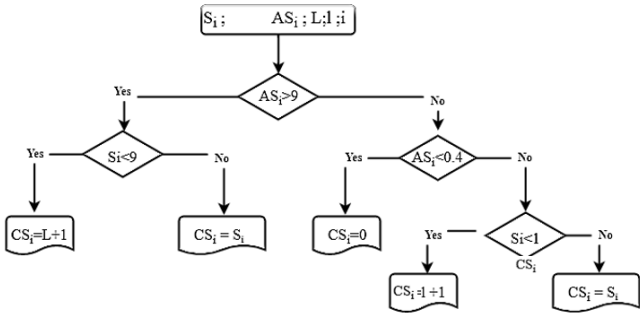


Fig. 5. The heuristic decision tree for correcting the speed

Figure 6 illustrates the speed correction obtained using the smoothing heuristic applied to the time-period of 280 seconds from 17400 to 17680 of a sweeping shift.

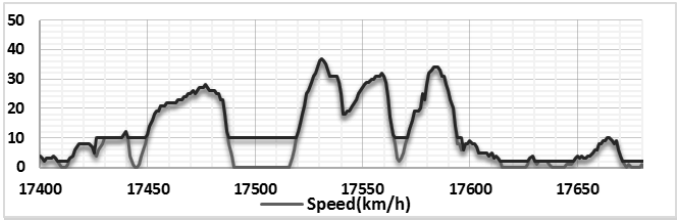


Fig. 6. Smoothing heuristic applied to the speed curve from 17400 to 17680 s

After cleaning and correcting the GPS database, the next step is to characterize the sweeping activity states based on the corrected speed value, since this attribute can be used for classification.

Manual Data Processing

About 17% of the data was processed manually through classification and clustering methods to infer the sweeping states and their attributes. Using the observed data, the broom states are identified based on the speed attribute. Let S be the speed broom in km/h. Then,

- If S in [0, 0.4[, the broom is assumed to be in the waiting state to account for the GPS speed error;
- If S in [0.4, 10[, the broom is assumed to be in the sweeping state;
- If S ≥10, the broom is assumed to be in the moving state.

The thresholds adopted are determined by estimation based on empirical findings.

State Characterizations

The state characterization is a multi-classification of records into various states (sweeping, waiting and moving). This is done according to the thresholds already described using MS Excel. Following this classification, all adjacent records classified into the same state are grouped together and their duration and average speed are calculated.

State Validation Using Video Data

After grouping the states according to the initial classification of the records, a validation of the grouping into states will take place based on the video images. Validation, in this case, serves two purposes. First, it helps verify that the identified states generated correspond to the real situation, otherwise, they are corrected. Second, states with a small duration are attached to adjacent states based on the video images.

Figure 7 illustrates the state characterization example according to the speed value for the period between 17000 and 18000 s.

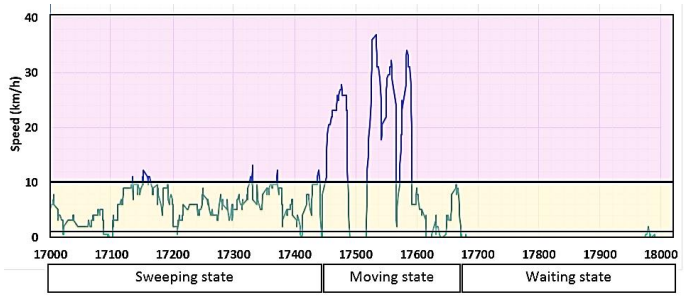


Fig. 7. States characterization according to the speed value validated by video images for the period 17000-18000 s

The validation of the classification based on the video images, as described above, takes a lot of time. This manual data processing was applied to the data from 11 shifts that were classified and validated using video images. Although this exercise proved valuable for understanding the processing of geolocation data, it is not practical for real-life applications given the big data involved. This motivated us to use the manual

processed data as an input to train powerful ML classification models. Our objective is to identify the best combination of training sets, input variables and classification algorithms that can be used to deal with the processing of geolocation data to infer the behavior of the sweeping system.

Machine Learning Model Building

While 17% of the data was processed manually, a framework similar to the manual process described above needs to be developed for the remaining data (83%). In practice, it is not possible to make a prediction of the GPS recordings to reach the sweeping states as defined above. FE is essential, in this case, to properly adapt the data for prediction algorithms learning considering the specificities of the GPS data. FE tools allow us to transform the manual procedure of sweeping state characterization into a learning procedure for sweeping behavior prediction. Thus, we will transform the phases of classification and correction into two phases of simultaneous classification-clustering.

Building the ML model based on FE involves four stages:

1. An initial classification of the records is performed based on the corrected speed. The manually processed shift data are used as training and testing sets for the ML classification models;
2. An initial state clustering based on the first classification is performed. In this stage, the duration and the average speed related to each state are computed;
3. A second classification based on the clusters resulting from the previous stage is carried out. The independent variables explored include the initial state, the average speed of the state and its duration;
4. A final state clustering based on the second classification is performed.

First Classification and Clustering

The first supervised classification is based on the corrected speed of the raw geolocation data as an input variable to classify the records into various states. Based on the training sets generated using the manual processing, the LR and the RF classifiers were used to achieve a correct classification. Python 3.7 is used to connect the library of classifiers and the assessment tools, including the confusion matrix and the accuracy indicator.

An initial clustering of states based on the first classification is performed. In this step, the duration and the average speed related to each cluster of states are calculated.

Second Classification and Clustering

The second classification was carried out using two methods. The first method used historical data to train the classifiers. However, the second method performed the training based on simulated data. A simplified simulation model was built based on the 11 manually validated shifts to generate the data instances used for training.

For the first method, the second classification is based on the cluster states generated in the previous clustering step where eleven data sets (shifts) were used for training and

testing. Different combinations of input variables (initial state, duration and average speed) were tested to obtain the best results. The two combinations that were retained include either all three variables or duration and average speed. Seven classification algorithms were used. Each of the 11 data sets is used in turn for training and then testing is done using the remaining ten data sets. This operation is performed with the two combinations of variables and uses all the seven classification algorithms leading to 1694 different results. These results are used to identify the best combination of training sets, input variables and classification algorithms in order to build the ML classification model.

For the second method, we have retained as input two variables (the duration and the speed of the state) since the initial state for the simulated shifts is missing. We also used the classifiers which gave the best results of the first method, namely LR and RF, Naive Bays (NB) and SVM algorithms. The selected classifiers were trained by each of the 10 simulated shifts and tested by each of the 11 historical shifts.

A final state clustering similar to the first one is performed based on the result of the second classification. This final step will produce a state database (observations) that provides the description of the sweeping system's behavior during a given shift.

Machine Learning Data Processing

The result obtained when applying building process described above allow the selection of the best classification model and the results of its application to one work shift.

To improve the classification prediction in the case of the last application, we used a corrective heuristic. This heuristic consists in eliminating the states with short duration (<60 s) and assigning them to the nearest adjacent state.

4 Results and Interpretations

4.1 Machine Learning Model Selection

The maximum accuracy allowed when we applied a unique classification of GPS raw data is 49.4%. This level of accuracy is considered insufficient, given the enormous number of misclassified states relative to the GPS data. For this reason, our idea is to carry out the classification in two steps rather than a single one using FE tools. Therefore, the second classification uses several input variables provided by the first classification which increases the prediction capability. This leads to a substantial improvement in the performance of the classification scheme. In this section, we discuss mainly the configuration of the ML classification model for the first and second classification.

First Classification

The first supervised classification is based on the corrected speed of the broom. The prediction of a certain shift proves that at least two classifiers perfectly performed this classification. These are the LR and the RF classifiers.

First Clustering

In this step, the duration and the average speed related to each cluster of states are calculated. This step is advantageous because it allows us to consolidate the second classification to remedy the defects of the classification based on the fixed thresholds.

Second Classification

Using real data

Table 1 presents the second classification results of the different combinations of input variables, classifiers and training set data, as explained in Section 3.2. Each entry in this table represents the average accuracy obtained using the corresponding shift as training set, the corresponding number of input variables and the corresponding classifier, where all the shifts are used for testing. The results show that the cases involving three independent input variables produce slightly better results than those involving two variables. Shift 11 and Shift 5 offer the best data sets for training in cases involving three and two input variables, respectively. The best classification algorithms are RF and LR in that order. The best result was obtained with Shift 11 as training set, three input variables and the RF algorithm with the accuracy of 86.56%. Table 2 provides the details of the best results in all shifts.

Table 1: The average accuracy of the alternatives explored

Classifiers	Variables	Training data sets										
		Shift 1	Shift 2	Shift 3	Shift 4	Shift 5	Shift 6	Shift 7	Shift 8	Shift 9	Shift 10	Shift 11
DT	2*	85.72	80.00	72.99	80.67	85.48	74.30	78.47	78.75	77.18	81.36	83.15
	3**	85.66	80.27	74.35	80.23	85.64	74.11	79.19	79.06	77.18	82.31	85.03
KNN	2	79.02	70.59	65.24	73.50	79.07	69.80	71.46	70.96	71.15	76.92	73.05
	3	79.12	71.24	65.51	74.66	79.07	70.39	72.09	71.76	72.16	77.41	74.34
KSVM	2	69.11	69.41	69.41	68.53	69.20	69.08	65.08	69.17	69.11	68.97	25.99
	3	69.11	69.41	69.41	68.49	69.20	69.08	65.07	69.17	69.11	68.97	25.96
LR	2	86.11	84.83	71.95	84.28	**86.22**	71.41	77.01	82.42	73.43	82.66	85.70
	3	86.09	84.75	72.80	84.34	86.37	71.98	77.35	85.20	73.97	83.82	85.92
NB	2	**84.87**	72.02	71.70	81.70	84.73	70.35	72.23	71.79	73.80	73.96	79.50
	3	86.09	84.75	72.80	84.34	86.37	71.98	77.35	85.20	73.97	83.82	85.92
RF	2	84.85	81.80	78.27	84.39	85.97	77.36	82.15	82.76	76.36	83.69	82.12
	3	86.05	81.47	78.50	81.45	**86.13**	78.69	82.76	81.95	76.17	85.40	**86.56**
SVM	2	85.87	79.93	70.85	86.09	**86.13**	71.38	77.10	78.17	73.10	85.06	85.66
	3	86.11	79.74	71.15	86.18	**86.36**	72.14	76.98	77.75	73.12	85.36	85.72

DT : Decision Tree; KNN : K- nearest neighbors; KSVM : Kernel SVM; LR : Logistic Regression; NB : Naïve Bayes; RF : Random Forest; SVM : Support Vector Machine

*: 2 variables - speed and duration of the state; **: 3 variables - initial state, speed and duration of the state

Table 2: The detailed results of the best solution (Training data: Shift 11)

Test data	**Shift 1**	**Shift 2**	**Shift 3**	**Shift 4**	**Shift 5**	**Shift 6**	**Shift 7**	**Shift 8**	**Shift 9**	**Shift 10**	**Shift 11**	**Average**
Accuracy	92.90	81.93	69.03	91.67	92.60	84.12	83.28	81.72	91.54	85.67	97.69	86.56

The two configurations retained for the ML classification models are illustrated in the Table 3.

Table 3: ML best classification models

Model	Classifier	Training set	Input variables	Average accuracy
Model 3V	RF	Shift 11	Duration, speed and initial state	86.56%
Model 2V	LR	Shift 5	Duration and speed	86.22%

Using simulated data

The manual processed geolocation data for 11 shifts were used to generate parameters for simulating the sweeping system. The simulation model developed was used to generate additional instances for 10 shifts that were used for training purposes.

When simulation data is used for training, only two input variables (speed and duration) were considered since the initial state variable was not available for the simulation data. Also, only the best four algorithms identified using real data were implemented.

Table 4 presents the second classification results obtained using simulated data as training data sets. In this case, we used the following classifiers namely LR, NB, RF and SVM. In this case, the best result is obtained using the 6th shift data set for training, all real data sets for testing and the LR classifier with accuracy 87.4%. Table 5 provides the details of this best result on all shifts. These results show that a better accuracy is obtained compared to the classification using real data sets.

Table 4: The average accuracy using LR, RF, SVM and NB classifier

classifier	S_shift1	S_shift2	S_shift3	S_shift4	S_shift5	S_shift6	S_shift7	S_shift8	S_shift9	S_shift10	Max	Avg	Rank
LR	87.34	87.02	87.27	83.75	86.73	**87.40**	87.17	87.05	87.21	87.16	87.40	86.81	1
RF	86.57	86.79	85.96	86.89	79.22	75.60	**86.92**	85.62	86.87	86.69	86.92	84.71	3
SVM	87.17	86.37	87.10	83.14	86.19	87.16	86.49	86.96	87.05	**87.22**	87.22	86.49	2
NB	61.65	54.59	73.60	53.73	**82.64**	58.12	57.36	60.35	75.14	54.49	82.64	63.17	4

Table 5: The detail results of the best solution (Training data: S_shift 6)

Test data	**Shift1**	**Shift2**	**Shift3**	**Shift4**	**Shift5**	**Shift6**	**Shift7**	**Shift8**	**Shift9**	**Shift10**	**Shift11**	**Shift12**	**Avg**
Accuracy	99.33	95.48	81.93	66.79	92.54	94.53	83.87	86.63	81.38	94.86	87.67	83.80	87.40

Second Clustering

A final state clustering is performed based on the result of the second classification. This final step will produce a description of the behavior of the sweeping system on a given shift. Although this second classification provides a categorization of states that is close to reality, some imperfections remain to be eliminated.

4.2 Results Comparison

To assess the prediction quality, the classification accuracy is used to compare the results obtained with real data to those obtained with simulated data as training sets. The prediction accuracy provided with LR and SVM classifiers, when trained with historical data, were in average 86.22% and 86.13%, respectively. These predictions using the same classifiers, when trained with simulated data, were in average 87.4% and 87.22%. Table 6 shows the comparison of this accuracy.

Table 6: Comparison of classification accuracy (%) between real and simulated training sets.

Classifier	Historical training data			Simulated training data		
	Max avg predict. (*)	Standard deviation	Rank	Max avg prediction	Standard deviation	Rank
LR	86.22	8.4	1	87.4	8.8	1
RF	85.97	7.6	3	86.92	9	3
SVM	86.13	8.4	2	87.22	8.9	2
NB	84.87	8.9	4	82.64	7.3	4

(*) : Average of all data test sets (11 shifts) related to the best shift trainer

FE tools have improved the performance of prediction algorithms in two ways. The first is to adopt two successive clustering-classifications in order to have more explanatory variables after the first classification and the second is to train the algorithms on simulated data. Figure 8 explains the results achieved using the adopted framework.

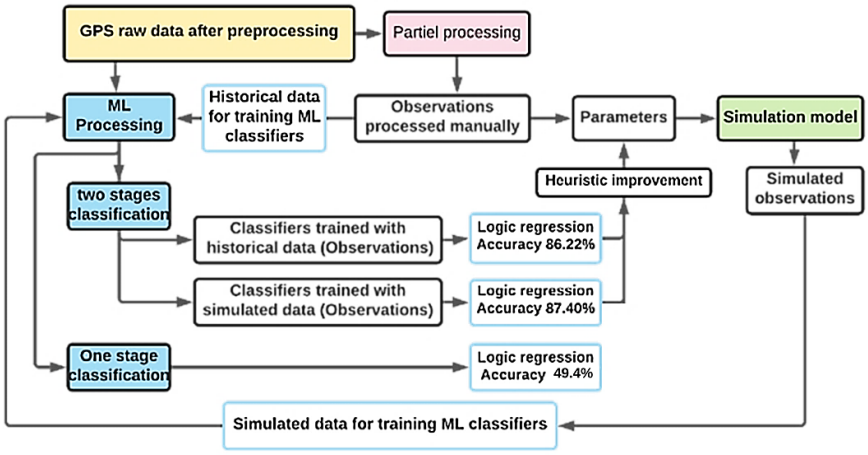


Fig. 8. FE Framework to improve the performance of prediction models

4.3 Corrective Heuristic Impact

In order to evaluate the impact of the corrective heuristic, described above, an application case is tested using a real shift and LR classifier trained with both historical and simulated data. Figure 9 illustrates the first and the second classification result for this particular application (States in the graph is noted 0: for waiting, 1: for sweeping and 2 for moving; the speed is in Km/h).

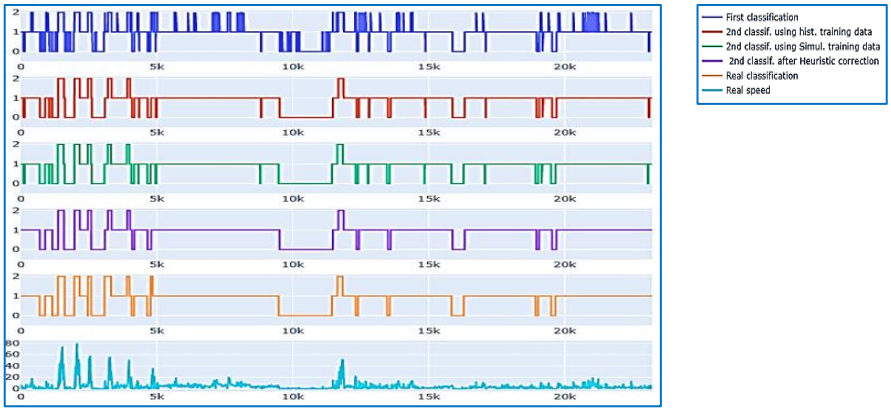


Fig. 9. Comparison of various classification schemes

The single classification has the accuracy of 49.4%. However, the second classification of the two classification model, when trained with historical data, has the accuracy of 95.48% on the training data. Its prediction accuracy is about 93.5%. However, when trained with simulated data, the accuracy on the training data is about 99.33% and its prediction accuracy is about 95.5%. When improved with the corrective heuristic, the classification trained with the historical data gives the accuracy of 98.4%.

The result of the second classification is a set of state observations as illustrated in the Table 7.

Table 7: State observations provided by the considered shift

Ref State	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	
Duration (min)	11.5	3.27	4.45	3.48	4.15	5.83	3.333	4.92	2	8.3	4.05	9.62	1.667	1.28	1.52	7.98	
Avg speed (km/h)	3.13	0.28	2.76	0.19	30.9	0.05	41.44	4.56	28	0.16	18.58	4.42	23.01	7.92	0.12	2.74	
States (*)	1	0	1	0	2	0	2	1	2	0	2	1	2	1	0	1	
Ref State	**17**	**18**	**19**	**20**	**21**	**22**	**23**	**24**	**25**	**26**	**27**	**28**	**29**	**30**	**31**	**32**	**33**
Duration (min)	2.15	1.23	77.2	32.7	3.37	3.17	7.967	1.43	18	1.68	37.88	7.47	44	1.53	7.97	2.6	59.2
Avg speed (km/h)	0.72	17.3	5.4	0.3	4.01	0.03	4.15	0.31	2.9	0.03	2.85	0.06	2.74	0.04	2.9	0	3.1
States	0	2	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1
(*) : 0 : waiting, 1 : sweeping, 2 : moving																	

Having the state observation database, we compute the simulation parameters that illustrate the sweeping behavior such as the state's frequency, the average speed and duration as shown in the Table 8. These parameters can be used for simulation if the database has a significant set of shift data.

Table 8: Parameters deduced from the shift observations

State	Number	Average duration (min)	Average speed (km/h)	Frequency (%)	Total duration (min)
0	13	5.8	0.00	39	75.2
1	14	21.1	3.83	42	295.2
2	6	2.7	26.54	18	16.4

4.4 Interpretations

The following observations can be made based on the results obtained:

- The ML tools can describe the behavior of the sweeping system with the accuracy of about 87%;
- Two-step classification-clustering scheme using FE tools improves the accuracy from 49.4% to 87.4%;
- The application of a corrective heuristic allows an improvement of the classification result of 5% when applied to a real shift data;
- Better results, in terms of accuracy and stability, are obtained with simulated training data compared to historical data.

5 Conclusion

This paper focuses on geolocation data processing to infer the behavior of a sweeping system in order to generate the necessary data needed to evaluate its operational and environmental performance using simulation and for potential AI-based applications.

A large range of classifiers, input variables, training and test data sets are used to build the multi-classification ML models. The results showed that LR and SVM are the best classifiers to process GPS data and some simulated instances constituted the best training data sets. The sweeping state prediction accuracy provided with LR and SVM classifiers when trained with historical data were in average 86.22% and 86.13%, respectively. These predictions using the same classifiers when trained with simulated data were in average 87.4% and 87.22%. These accuracy predictions are stable when the classifiers are trained with simulated data.

The main contribution of this paper is the use of the FE tools to transform a manual classification into a classification driven by machine learning algorithms. For predicting the behavior of a logistics system from GPS data, a double classification allows prediction based on attributes other than speed. Compared to the prediction accuracy reported in the literature, as discussed in Section 2, the accuracy levels presented in this research are very promising. Such an improvement was possible using a corrective heuristic that enhanced the classification of the sweeping system states. The ability to infer the behavior of the sweeping system, based on geolocation data processing, will form the basis for planning this type of operation in the future, for improving its sustainability and for its possible virtualization based on AI applications within the framework of a smart city.

Acknowledgments

The work in this paper was funded by Fonds de Recherche du Québec - nature et technologies (FRQnet), grant 2019-GS-260551, in partnership with street sweepings service provider (Arseno Sweeping). These supports are gratefully acknowledged.

References

1. Kim, B. S., Kang, B. G., Choi, S. H. & Kim, T. G. Data modeling versus simulation modeling in the big data era: case study of a greenhouse control system. *SIMULATION* **93**, 579–594 (2017).
2. Laranjeiro, P. F. *et al.* Using GPS data to explore speed patterns and temporal fluctuations in urban logistics: The case of São Paulo, Brazil. *Journal of Transport Geography* **76**, 114–129 (2019).
3. Servos, N., Liu, X., Teucke, M. & Freitag, M. Travel Time Prediction in a Multimodal Freight Transport Relation Using Machine Learning Algorithms. *Logistics* **4**, 1 (2020).
4. Gong, L., Kanamori, R. & Yamamoto, T. Data selection in machine learning for identifying trip purposes and travel modes from longitudinal GPS data collection lasting for seasons. *Travel Behaviour and Society* **11**, 131–140 (2018).
5. Pluvinet, P., Gonzalez-Feliu, J. & Ambrosini, C. GPS Data Analysis for Understanding Urban Goods Movement. *Procedia - Social and Behavioral Sciences* **39**, 450–462 (2012).
6. Shen, L. & Stopher, P. R. Review of GPS Travel Survey and GPS Data-Processing Methods. *Transport Reviews* **34**, 316–334 (2014).
7. Govindan, K., Cheng, T. C. E., Mishra, N. & Shukla, N. Big data analytics and application for logistics and supply chain management. *Transportation Research Part E: Logistics and Transportation Review* **114**, 343–349 (2018).
8. Khurana, U., Samulowitz, H. & Turaga, D. Feature Engineering for Predictive Modeling Using Reinforcement Learning. *Proceedings of the AAAI Conference on Artificial Intelligence* **32**, (2018).
9. Khurana, U., Turaga, D., Samulowitz, H. & Parthasrathy, S. Cognito: Automated Feature Engineering for Supervised Learning. *2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW)* (2016) doi:10.1109/ICDMW.2016.0190.
10. Nargesian, F., Samulowitz, H., Khurana, U., Khalil, E. B. & Turaga, D. Learning Feature Engineering for Classification. *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence Main track.* 2529–2535 (2017) doi:https://doi.org/10.24963/ijcai.2017/352.
11. Montini, L., Rieser-Schüssler, N., Horni, A. & Axhausen, K. W. Trip Purpose Identification from GPS Tracks: *Transportation Research Record* (2014) doi:10.3141/2405-03.
12. Ribeiro, J., Fontes, T., Soares, C. & Borges, J. L. Process discovery on geolocation data. *Transportation Research Procedia* **47**, 139–146 (2020).
13. Dabiri, S. & Heaslip, K. Inferring transportation modes from GPS trajectories using a convolutional neural network. *Transportation Research Part C: Emerging Technologies* **86**, 360–371 (2018).
14. Guo, S. *et al.* GPS trajectory data segmentation based on probabilistic logic. *International Journal of Approximate Reasoning* **103**, 227–247 (2018).
15. Feng, T. & Timmermans, H. Comparison of advanced imputation algorithms for detection of transportation mode and activity episode using GPS data. *Transportation Planning and Technology* **39**, 1–15 (2016).
16. Bhavsar, P., Safro, I., Bouaynaya, N., Polikar, R. & Dera, D. Machine Learning in Transportation Data Analytics. in *Data Analytics for Intelligent Transportation Systems* 283–307 (2017). doi:10.1016/B978-0-12-809715-1.00012-2.