European Transport \ Trasporti Europei (2021) Issue 83, Paper n° 5, ISSN 1825-3997

https://doi.org/10.48295/ET.2021.83.5

Support vector machines and k-means to build implementation areas of bundling hubs

Jihane El Ouadi[1,2,3,*], Nicolas Malhene[3], Siham Benhadou[1,2], Hicham Medromi[1,2]

[1]*Research Foundation for Development and Innovation in Science and Engineering, Casablanca, 8118, Morocco*
[2]*National and High School of Electricity and Mechanic, HASSAN II University, 8118, Casablanca, Morocco*
[3]*EIGSI, La Rochelle-Casablanca, 17041-20410, French-Morocco*

Abstract

City transportation has three basic components that create the essential environment for its functioning and the social welfare namely infrastructure, operational assets, and management policies. The key focus of this article is on understanding long-term distribution of transport demand in order to build bundling networks. To achieve this aim, we provide a hybrid machine-learning approach using a combination of several clustering and forecasting algorithms that are considered efficient given the key performance indicators obtained. This approach involves combining two types of algorithms: clustering and prediction algorithms. Based on simulated benchmarks, results indicated that the clustering phase is still appropriate using the k-means algorithm. To improve the k-means results, we measured 30 validation indices to estimate the number of clusters. In so doing, not only does it want to validate the clusters but also to identify the optimal k. To evaluate forecast accuracy in the demand prediction phase, we used the standard key performance indicators, namely MSE, RMSE, MAPE and R^2. The SVM algorithm has been judged as the most efficient prediction algorithm based on average values of the obtained metrics.

Keywords: Hub location problem; freight bundling; transportation demand; urban zoning; machine-learning.

1. Introduction

City transportation has three basic components that create the essential environment for its functioning and the social welfare namely infrastructure, operational assets, and management policies. Logistics and transportation are an important part of these components, which impact per capita comfort and reception in well-established patterns. Until recently, travel demand has tended to increase significantly with economic and population growth. The use of roads remains predominant for transporting both passengers and goods. Comparatively speaking, the urban logistics infrastructure that

represents the supply side, particularly in emerging countries, has lagged far behind demand. These trends continue in cities with restricted land area and facilities, but with an increasing number of journeys and deliveries being made daily.

The propensity of cities to have congested road networks challenges the key logistics issues of traffic flow and accessibility. In other words, the irrational exploitation of logistics resources to serve a limited number of customers leads to vital goals of achieving efficient logistics performance and optimal land use. In this regard, the city could be relieved of some unnecessary freight movements by public passenger transport such as buses, streetcars and subways (El Ouadi et al., 2021). The exploration of these mass transportation systems to meet supply needs is consequently supposed by the residual transportation capacity that remains during off-peak hours.

Shared transport systems remain fairly common in long-distance distribution as part of first-mile delivery. However, the benefits in the last few miles seem undeniable, making it an attractive future alternative. Integrating the two flows allows users to leverage the remaining capacity of passenger networks to move freight from the flow generators to the end customer. With the cargo bundling, the use of the remaining capacity will become more efficient and less expensive. However, when considering quality and comfort, the attractiveness of the system should be determined in terms of city data available, real-time information and connected systems. As cities become increasingly intelligent, artificial intelligence systems can build confidence in these patterns, thereby improving sustainability.

In general, urban travel data shows seasonal fluctuations due to expected and unexpected events (vacations, national and religious celebrations, …). In this paper, we focus on assessing demand to reduce uncertainty, which may involve freight bundling, i.e., a shift in focus. Exploring decision support in an intelligent framework could have a positive impact on the process being addressed. Therefore, we performed accuracy tests of machine-learning algorithms for an effective demand study tool that helps in the long-term implementation of mobility systems and networks.

The rest of the paper is organized as follows: Section 2 clarifies the context in which the methodology is applied. Section 3 presents a brief state of the art of zoning approaches that are involved in urban zoning setting. Section 4 explains the proposed approach. Section 5 of the paper is devoted to the implementation and the results of simulations. Accuracy-based benchmarks will be presented in this section. Last but not least, section 6 concludes by proposing further research directions.

2. Context of application

2.1.Problem statement

Rationalizing the use of logistics patterns, equipment, and facilities, in general, is worth exploring the validity of the decision in a long-term, increasingly big data future. As with any integrative logistics plan, shared transportation amplifies risk by broadening stakeholder participation in the decision-making processes. From a technical point of view, it represents a highly competitive alternative where any wrong decision can have irreversible impacts. In order to change the direction of the flow while minimizing risk and improving the efficiency of the shared transport system, goods will be stored in an urban bundling center (hub) before being reloaded into another vehicle with sufficient remaining capacity.

A two-tiered bundling network allows deliveries and part-loads to be grouped into collective shipments. The first tier is dedicated to logistics sites located ahead of the public transport network. Its purpose is to receive flows that must join the urban area and to coordinate their shipment by public transport. The second tier is made up of urban hubs that are located along the mass transport network. They are responsible for collecting the supplies and organizing the movements to the final destinations. Similarly, in order to cover the final recipients with delivery routes, light distribution vehicles must be assigned to each hub (Figure 1).

Figure 1: Shared Public Transport in terms of passengers and freight flows.

Since the mass network is usually permanent, special attention should be devoted to the second-tier structure, i.e., the location of the hubs in the city. Therefore, the potential location of hub facilities should attract consumers with full consideration of the geographic concentration where demand for freight transportation is high. Similarly, the capacity of hubs and their capillary fleets should be significantly tailored to the geographic areas hosting demand nodes in terms of local businesses, enterprises, storage points, etc. (Zaho et al., 2018). Indeed, the study of logistics demand referring to a given quantity of cargo to be shifted from A node to B node can help in successfully locating such hubs. In addition, this demand study will be useful to size light fleets in order to avoid oversizing the logistics resources. Thus, the main objective is to reduce the uncertainty of the logistics demand that leads to additional costs due to the oversizing of the system. In other words, we seek, through a zoning approach, to avoid the long-term implementation problems of logistics systems that occur when certain forecast measures (i.e., the demand to be shipped) are not taken into account in advance.

Table 1: Types of bundling hubs versus the size of implementation zones.

Logistic facilities	*Size*	*Ease of implementation*
Local logistic facilities (LLF)	Approx. 500 m² (5382 sq. ft)	High
Urban Distribution Centers (UDC)	Approx. 1000 m² (10764 sq. ft)	Low
Urban Logistic Zone (ULZ)	4 hectares	Medium

2.2.Hub location problem

One path in literature focused on locating urban facilities as being derived from spatial zoning. Spatial zoning is defined as the appropriate preliminary phase addressing strategic bundling layout as more ways to reduce the average length of haul and total traveled vehicle kilometers (Wygonik and Goodchild, 2018), (Janjevic and Nadiaye, 2014).

Depending on the size of the covered zone as depicted in (Table 1), a bundling facility is likely to be Local Logistic Facility (LLF), Urban Distribution Center (UDC) or Urban Logistic Zone (ULZ).

In order to propose a rational, sustainable, and efficient policy, transportation research is focusing on alternatives for pooling resources. One promising alternative is to make rational use of existing facilities on the basis of residual transport capacity, particularly public passenger transport. In many urban areas, there is a complementarity between passenger and freight movements. Nevertheless, the two flows may compete for integration policies from limited available land and transport infrastructure. The delivery of freight is, in fact, carried out by large vehicles owned by suburban logistics providers, to be delivered in identical directions. These systems are based on the concept of shared vehicles (commuter/cargo) with an emphasis on capacity, not vehicle type. It is an enhancement of the existing system that will not add additional traffic but could reduce the traffic that carries freight. Aiming for more coverage and less unnecessary trip generation, the proposed system should typically rely on bundling hubs.

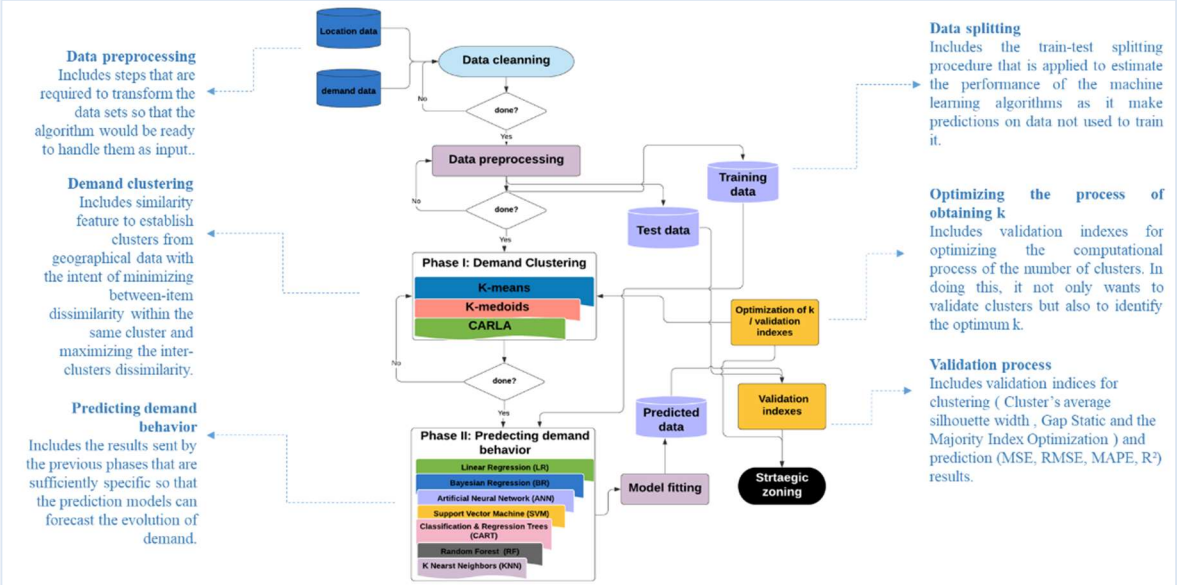


Figure 2: Schematic flowchart of the proposed approach.

The supply chain leads to a set of suppliers, businesses, retailers, customers, and facilities. The location of the BH is a key decision that should impact the entire supply chain management. It represents an imperative managerial task that can be achieved through trade-offs between demand, capital, and coverage levels. Like any network design problem (a review is provided here (Iliopoulou et al., 2019)), the problem related to the aggregation process is divided into three hierarchical layers: the strategic, tactical and operational layers (Segura et al., 2019). The strategic layer evaluates the long-term planning of the network, including the analysis of facility location problem. This is the highest level of planning that involves long-term investments once the distribution network is established.

In contrast, the operational layer deals with shipments and routes to improve the daily flow of goods. Although it has some similarities to the operational layer, the tactical layer

is generally short-term and recovers the results and constraints of the operational layer (fleets, drivers, facilities, legal constraints, etc.).

For bundling networks, in particular the hub location problem, is the key to strategic decision-making. Given the context, the chosen configuration must allow to capture as much demand as possible over the long term without having to modify or even update the built patterns. In other words, bundling hubs present a permanent risk that it is irrational to disclose in the short term. This means that a bad location could result in damage based on the following factors: cost, environment, and social risks.

In light of these facts, selecting the appropriate hubs could be a challenge in real-world systems, especially in the case of supply shortages. The main concern of this paper is to provide a long-term splitting of city to support hub location analysis by realizing zoning approach using machine-learning.

Since the level of customer coverage is influenced by the amount of demand and delivery behaviors, the inclusion of the demand forecasting process could reduce the uncertainty in the location selection. Next, the first step of clustering is a macroscopic analysis of the demand distribution to assess the potential of each territory to host a hub. Then, each built demand area is evaluated in themes of demand evolution in the future with a forecasting step. In this respect, the effectiveness of the developed zoning algorithm is demonstrated through benchmarks. From a practical point of view, the empirical results show that the proposed solution methodology has provided valuable insights relying of popular metrics.

3. Existing approaches

Splitting urban area has been addressed in (Skai, 2016a) as a means of segmenting logistics hub location to reduce overall truck travel. Actually, this paper has developed discrete models to evaluate land use based on the specific needs of the target districts of the facilities. Based on these models, they showed that the effect of zoning policies is infinitesimally major to generate economies of scale in shipping distances. The author of the article (Rulence, 2003) has considered the problem of opening and closing commercial establishments as a function of territorial coverage as well as spatial coherence. In this respect, relative entropy is applied to estimate the concentration of outlets. It is used in this research to evaluate the density of units in each cell in an attempt to understand how facilities are distributed and how homogeneous the coverage of the areas is. Moran's autocorrelation coefficient, which is built on adjacency matrices, tested for spatial coherence in order to reveal whether there is a relationship between the proximity of places and their level of similarity. This study has treated areas as equivalent, even though the distribution within each area may be different.

A similar principle has been also used to select appropriate locations for rail terminals integrated in a European hub-and-spoke network (Limbourg et Jourquin, 2007). The basic idea has been to use freight streams and their spatial patterns as a starting point for locating hubs within a median p-hub problem. In the paper (Huang et al., 2018), they established a model-based cell block to be served by sets of vehicles. Increasing the number of cells, for example, requires an increase in the number of satellites, which may not be easy in densely populated urban areas where land is typically expensive and not readily available. Therefore, the authors considered an extended BDP (Block Design Problem) concept that provides geographic regions served by small groups of drivers in a two-echelon network distribution problem. As a result, they showed that the constructed blocks allow a two-echelon delivery system to handle daily variations in delivery volumes

in a cost-effective manner. (Ducret et al., 2016) have presented a conceptual model dedicated to the evaluation of the urban territory by empirical investigation to optimize the last mile of urban logistics. It consists of two steps that link urban freight to urban form: a territorial diagnosis based on clustering analysis and module recommendations for each type of area. In the article (Baro et al., 2016), they proposed a zoning method based on a radio-centric grid that focuses on urban areas, taking into account population densities and built-up areas. This zoning is adopted for evaluating the structural role of transportation networks and for characterizing peripheral areas that are sometimes poorly discriminated in transportation studies. Nevertheless, this approach can provide a centrality of the city and then it can aggravate the congestion problems. The authors of (Noves et al., 2009), (Galvao et al., 2006) has focused on zoning models using the Voronoi method to address the problem of continuous district location. They dealt with smooth district contours starting from a previously determined circular radial in order to perform urban zoning using socio-economic indicators. The paper (Carlsson and Devulapolli, 2013) has defined a multiplicatively weighted Voronoi diagram to solve the problem of dividing, based on proximity, a geographic region into subregions so as to minimize the maximum workload of a set of facilities in that region. However, this model requires that all subregions have the same area.

Thiess polygons have been used in the paper (Kazemzadeh-Zow et al., 2017) to identify districts with different morphology and attributes. The regularity of polygon boundaries may prevent an accurate mapping of polygons to district boundaries. In addition, continuous urban growth may affect the stability of the polygon by adding or removing some features. Shannon entropy has been presented in (Delaitre et al., 2008) as a conceptual tool that evaluates the distinct homogeneity in terms of the scatterplot of multiple quantified variables regarding the positions of "individuals" in the city. The authors of (Manganelli and Murgante, 2012) have presented a zoning scheme that split the town into socially and environmentally homogeneous sub-areas. This model used statistical data and neural networks to develop territorial data sets. The research cited here (Austin et al., 2008) processed geocoded address data and, by quantifying the median distance using bivariate statistics, examined the clustering of fast-food restaurants.

4. Adopted approach

The articles mentioned above reveal that the development of strategic urban zoning schemes for demand assessment is already being addressed for various localization purposes. Nevertheless, it has not been addressed in a machine-learning framework, even though it could provide a new and easy-to-use framework capable of providing better accuracy than conceptual approaches. The objective of this paper is to endorse the findings of our previous papers (El Ouadi et al., 2020) for validating the efficiency of the proposed approach.

The proposed approach consists of three main steps: data pre-processing, demand clustering, and finally demand prediction. The sequence of steps and the links between them are described in (Figure 2).

4.1.Data pre-processing

Data is a known key that has the greatest influence on the large-scale accuracy of machine-learning algorithms. Thereafter, due to the lack of recorded information on the manner, location and number of goods delivered in the city center, the most appropriate way to obtain data is to conduct a field investigation. Such a field investigation is designed

to collect data in the form of attributes that contain data on the company ID, name, contact information, geographic location, type of operation (pickup, delivery), date, etc. Nevertheless, for full credibility, in this paper, the datasets have been sourced from two web pages, the links to which are provided here (https://opendata.vancouver.ca/, http://www.platinum.matthey.com/).

In order to ensure meaningful benchmarks, preliminary data processing is performed to reduce errors on both processes and improve the learning speed (Figure 3):

i) Data Scaling: In this step, we intended to perform both cleaning of data from outliers, overlaps and gaps. In this regard, we developed the data normalization for the convergence performance. The main concern is to cope with the Min-Max method that takes advantage of the current distribution of variables.

ii) Clustering tendency: an important limitation arises when clustering data, as the associated algorithms may still return clusters even if the data do not return clusterable sets. To avoid this problem, we examined the clustering ability of the data upstream using the ordered dissimilarity image algorithm.

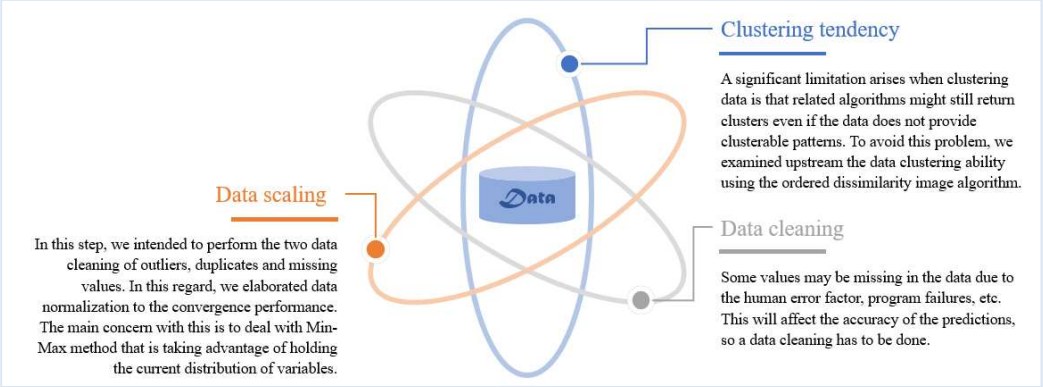


Figure 3: Data pre-treatment.

Ideally, the training of the hybrid algorithm in this approach starts with the preparation of raw data and its adaptation to the following phases of clustering and forecasting:

4.2.Clustering models

The clustering step identifies which geographic areas are hosting a shipping and/or pickup demand and which zones are not. In general, a clustering algorithm includes a similarity function that seeks to establish k -clusters from geographic data with the goal of minimizing dissimilarity between items within a cluster and maximizing dissimilarity between clusters. According to the reviewed literature, urban areas are different from suburban areas in that they involve "proximity" as a factor that may affect the distribution of goods (Ros-McDonnell et al., 2018). Therefore, we seek to tackle an optimization problem with a distance-based correlation feature. Such a clustering problem implies that two items are similar if there is a strong correlation between their features. Starting from the distance matrix construction, both Euclid and Manhattan distances are tested in this step to calculate the degree of similarity of two geographical data.

To achieve this goal, logistic demand is a concept that explicitly emphasizes the construction of clearly distinct areas. Thus, the numerical nature of the data prompted us to implement the k-Means and k-medoids clustering algorithms that are widely considered practical and effective for geocoded data (Xu et al., 2018). However, much

research has been conducted to obtain a faster algorithm that is applicable to large-scale datasets. Instead of computing medoids for each dataset, we tested clustering using the CLARA algorithm with fixed sample size data, so that an optimal set of medoids is generated for each sample (Appendix 1).

4.3.Forecasting models

How does demand behave in a given area? The answer to this question will determine whether or not the demand for transportation in that zone is certain in the long-term. Considering the dynamics of demand, zones could respond on various scales: zone limits and growth in zone capacity. Generally, bundling hubs remain static installations that involve a financial investment requiring long-term amortization that is incompatible with demand dynamics. This remains a powerful enough constraint to highlight these dynamics rather than evaluate the growth in zone capacity.

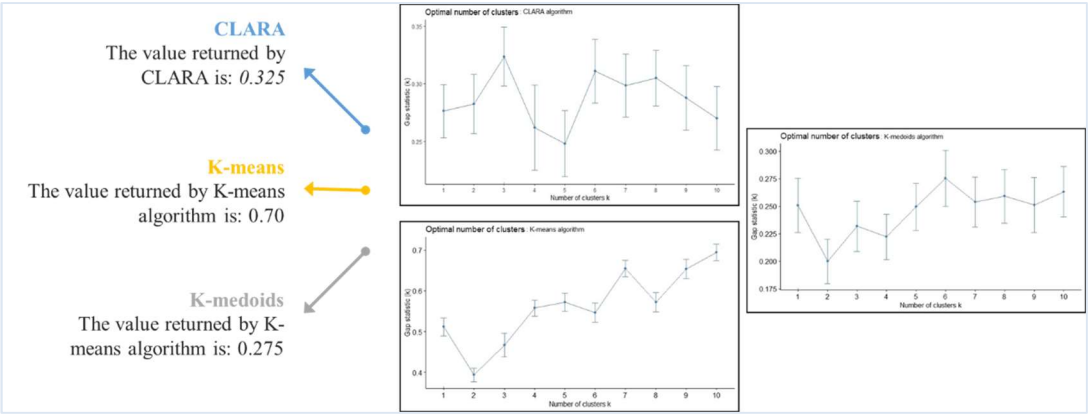


Figure 4: Determining the number of clusters k.

Both qualitative and quantitative methods provide the main forecasting methods in the literature.

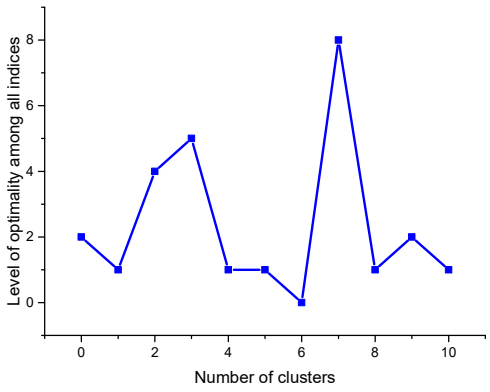


Figure 5: Optimal number of clusters *k* according to 30 indices.

Qualitative techniques consist of questionnaires or field surveys that address cases of unavailable data. These techniques are sometimes expensive and their predictive results are not always adequate. Quantitative forecasting techniques apply time series based on recognition, modeling, and patterns in historical datasets. Historical, statistical and machine-learning approaches are the most important classes of quantitative methods. For

the purpose of our case, we intend to apply machine-learning based techniques for higher forecasting accuracy and update option (Martin et al., 2019).

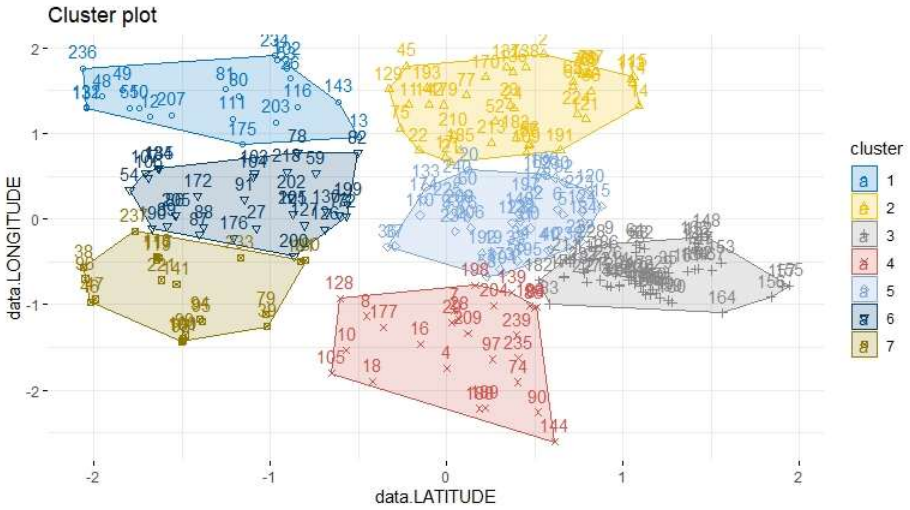


Figure 6: A map capture of built clusters.

These techniques provide an added benefit of learning and training from the historical input data and continuously adapting the forecasts with the input data. Without loss of generality, time series methods such as moving average, exponential smoothing, and autoregressive integrated moving average models forecast future demand based on prehistoric data. Statistical methods such as regression-based models or causal methods rely on the assumption that the demand forecast is related to a set of factors through maximizing the correlation between them. Machine-learning methods, such as Artificial Neural Networks (ANNs) and k-Nearest Neighbor Networks (k-NNs), are used for classification and regression, while support vector machines (SVMs), decision trees, and Adaboost classifiers improve the predictive ability with sample data. Model-based approach tools (e.g., Kalman filtering) rely on recursive mathematical equations to minimize the squared error, thus predicting short-term demand (Yu et al., 2018). The models considered robust are tested in this section. For an overview of the models, the reader might consider (El Ouadi et al., 2020), (Yu et al., 2018), (Bianco and Nardini, 2013), (Wang et al., 2019), (Xin et al., 2016) (Figure 10 in Appendix 2).

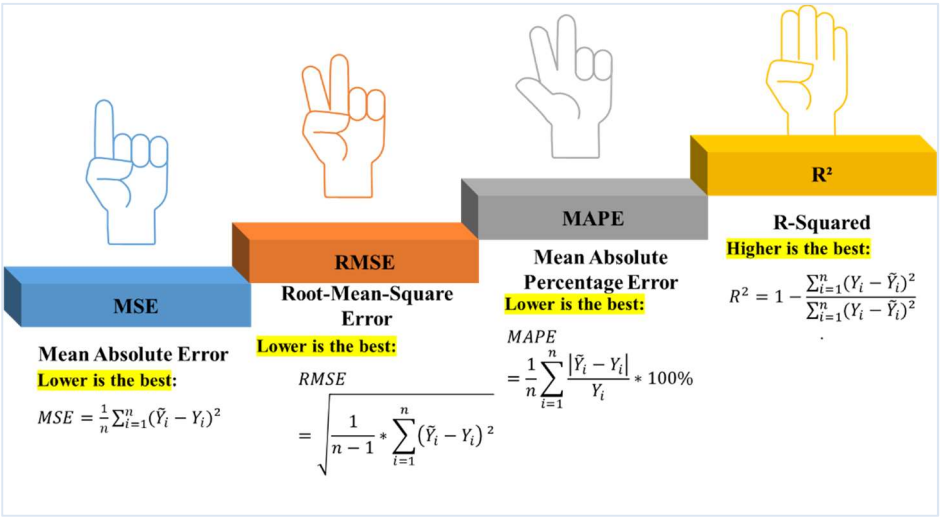
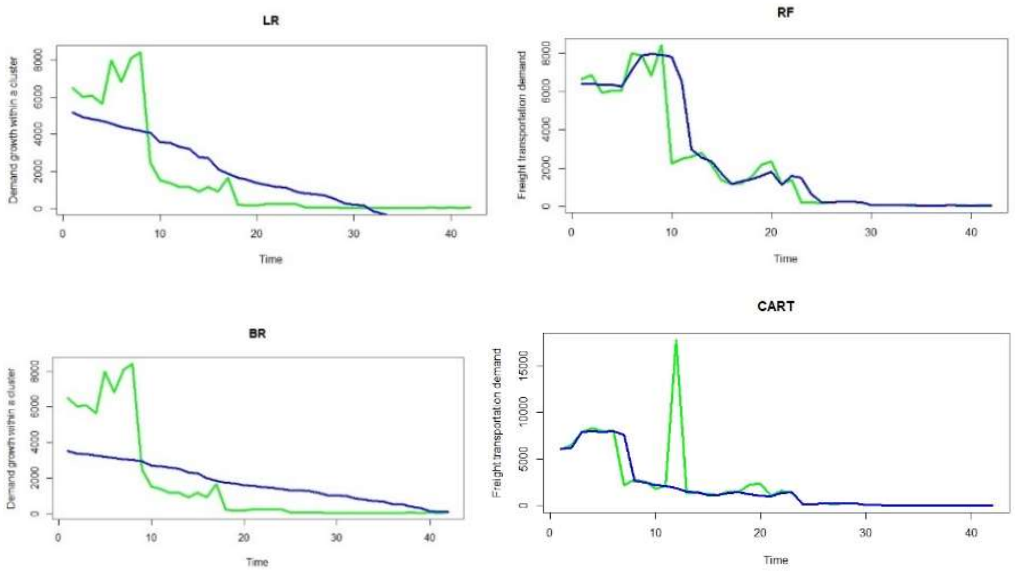


Figure 7: Key performance indicators of forecasting algorithms.

5. Implementation and results

The main issue that impacts the validity of the clustering results is the number of clusters that will be constructed.



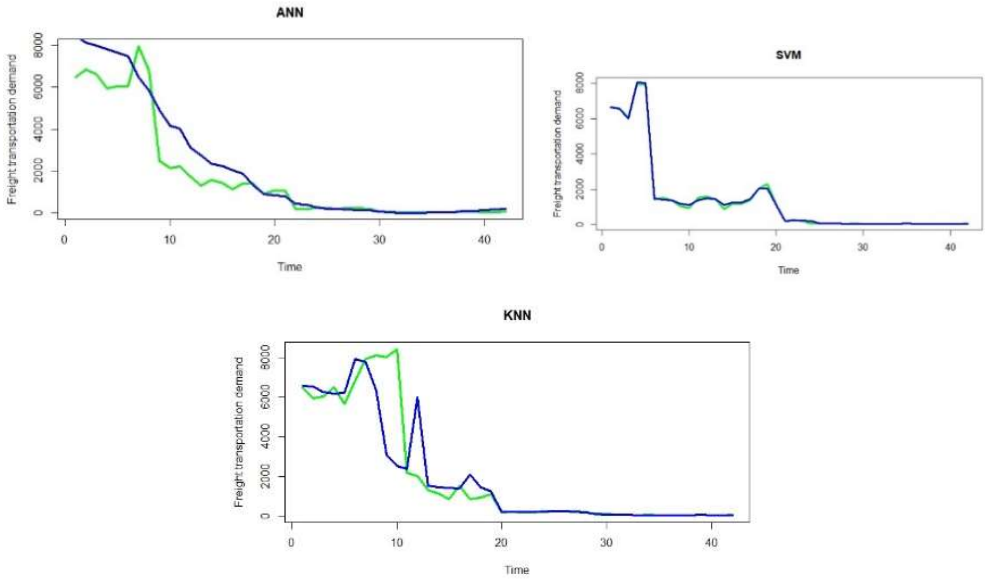


Figure 8: Logistics demand prediction: SVM model achieved the better performance.

The reviewed literature indicates that there is no consistent standard for determining the optimal number of clusters. However, this task may depend on the optimization of the indices that are involved in clustering validation tests. Clustering validation methods fall into two important classes: internal and external methods. Internal methods measure the similarity between nodes in the same cluster. External methods check the dissimilarity between clusters. In this regard, we evaluated the overall intra-cluster variability over several k-values against their values that are estimated to be below the null distribution of a statistic under the null hypothesis. Thus, the optimal cluster size is a value that maximizes the variance statistic, which means that the clustering process does not come from the random uniform distribution.

The results indicate that the optimal number of clusters varies from one algorithm to another while obtaining the maximum overall gap statistic (Figure 4). Therefore, the results showed that this type of clustering is further improved by using the k-means algorithm. The reason for the poor performance of k-medoids and CLARA is due to the fact that their processes always consider low static gap values (=0.275 and 0.325, respectively) compared to that of k-means (=0.70). Due to the data pre-processing and optimization process, the results of K-means, CLARA, and k-medoids did not result in clusters with a negative silhouette or oversized clusters (Table 3).

To further improve the k-means results, we evaluated 30 validation indices to estimate the number of clusters using the NbClust package. The aim of this process is to take the best clustering scheme among the different results obtained by varying any possible pattern of clusters, cost function and validation methods. In doing so, it not only validates the clusters but also identifies the optimal k. Following the rule of the majority, the best number of clusters is 7 (Figure 6) while getting the maximum overall clustering validity index (Figure 5).

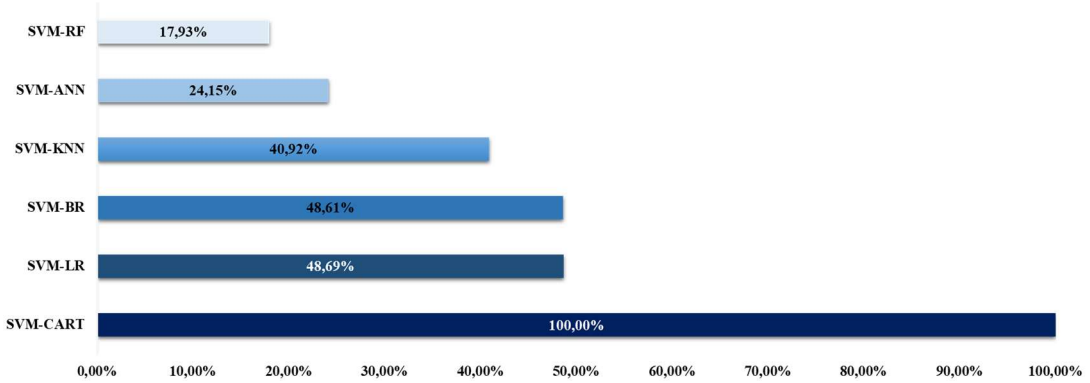


Figure 9: SVM algorithm is still the most efficient based upon average values of the tree indicators of performance (the R^2 metric is close to 100%).

Expanded on a simulated system, it is interesting to note that without understanding the stability of the clustering results, no optimal results could be performed for such geocoded data. In this regard, another evaluation of the consistency of the clustering result based on the cross-classification method is performed on the complete data.

These validation measures fall into three general categories: internal, stability and biological. The internal phase includes measures of the internal and external cluster validation indices. The stability of results is based on the average non-overlap proportion (APN), the average distance (AD), the average distance between means (ADM) and the figure of merit (FOM). Finally, the "biological" phase provides two biological validation measures, the Biological Homogeneity Index (BHI) and the Biological Stability Index (BSI). The average of the overall indices validates the above results. This stability test has been performed by the function clValid().

Since the objective of this paper is to provide the framework of an approach rather than a solution for a specific case study, we aim to illustrate the hypothesis of growth in mobility demand assuming that forecasts are made inside a cluster.

For evaluating forecast accuracy and benchmarking the selected models, we selected the key performance indicators, namely MSE (Mean Square Error), RMSE (Root Mean Square Error), MAPE (Mean Absolute Percent Error) and R-squared (R^2) (Figure 7). For all the prediction models studied, we have kept the same sharing rates in order to provide an appropriate benchmark: 70% is used for training and 30% is dedicated to testing model performances. Similarly, in order to perform this sampling, we have implemented a k-fold cross-validation that allows us to test the efficiency of several fitted sets. The graphs below show the results obtained from the logistic demand forecast with the ensembles obtained from the data test (Figure 8).

According to the obtained MSE and RMSE performance measures (Table 2), we could state that SVM performs 17.93% better than RF, 24.15% than ANN and 40.92% than KNN. Moreover, we can see that it fits the data better than BR with a rate of 48.61%. But compared to LR and CART, we notice that the accuracy rate of SVM reaches 48.69% and 100%, respectively. For the remaining MAPE metric, it is shown that SVM outperforms the RF model with a score of about 6.59%. Comparing this model with ANN and KNN, it remains 5.96% and 1.64% better respectively. The BR is less accurate in predicting logistic demand with a score of 93.05%. Nevertheless, CART and LR are equally less reliable with ratios around 0.63% and 100%. Still judging the SVM as the most efficient

based on the average values of the tree performance indicators, we can see that the R^2 metric is close to 100% (Figure 9).

Table 2: Performances of the tested forecasting algorithms.

	LR	*BR*	*ANN*	*KNN*	*CART*	*SVM*	*RF*
MSE (x e+06)	2.5216 90	2.5189 47	0.9115 076	1.972 475	6.7568 61	0.0059 53015	0.5838 131
RMSE (x e+06)	0.0015 87983	0.0015 87119	0.0005 47291	0.001 40444 8	0.0025 99396	0.0000 77155 78	0.0007 64076 6
MAPE	8.0028 66	7.5019 06	0.5976 764	0.255 2873	0.1775 2	0.1243 383	0.6455 585
R^2	0.6421 846	0.6431 846	0.9230 127	0.762 2745	0.4505 758	0.9987 86	0.7268 342

Table 3: cluster's average silhouette width.

	K-means		*K-medoids*		*CLARA*	
	Size	Ave.sil. width	Size	Ave.sil. width	Size	Ave.sil. width
1	40	0.43	32	0.39	13	0.53
2	14	0.38	41	0.46	22	0.55
3	30	0.38	53	0.54	11	0.40
4	19	0.58	26	0.36	-	-
5	30	0.35	46	0.38	-	-
6	23	0.28	42	0.37	-	-
7	15	0.38	-	-	-	-
8	17	0.49	-	-	-	-
9	20	0.34	-	-	-	-
10	32	0.37	-	-	-	-

Appendix 1

Formally, our clustering problem aims to solve an optimization problem of (*), where ∂_K is the centroid of the cluster C_k C_k and $\sum_{LP}\|LP_i - \partial_k\|^2$ is the variance (El Ouadi et al., 2020).

$$Min_{c_1,..,c_k;\partial_1,....,\partial_k} \sum_{j=1}^{k} \sum_{i \in c_j} \left\| LP_i - \partial_j \right\|^2 (*)$$

K-means algorithm

1. Selecting $\partial_1, \partial_2, .., \partial_k$ randomly from $LP = \{LP_1, LP_2, \dots, LP_n\}$;
2. Determining the cost function (Euclidian/Manhattan distance-based correlation) between each location LP_i and centroid ∂_j relying on distances matrix:
3. $dji = \|LP_i - \partial_j\|^2 ; 1 \leq j \leq k \; and \; 1 \leq i \leq n$
4. Updating clusters: assign each LP_i to the nearest centroid ∂_j ;
5. Updating centroids once the PLs are assigned;
6. Computing the objective function:

7. If converges (*): thus, cluster centers do not differ from the preceding iteration as the algorithm generates the final cluster centers.

8. Else: steps 2 to 5 are repeated until the objective function converges.

K-medoids algorithm

1. Select k random points out of the n data points as the initialization medoids.
2. Associate each data point to the closest medoid by using the cost function.
3. While the cost decreases:
4. For each medoid ∂_j, for each data LP_i point which is not a medoid:
5. Swap LP_i and ∂_j, associate each data point to the closest medoid, recompute the cost.
6. If the total cost is more than that in the previous step, undo the swap.

CLARA algorithm

1. Randomly build several subsets of constant sampsize from the original dataset
2. Run the PAM algorithm on each subset and choose the corresponding k corresponding medoids.
3. Attribute each observation LP_i of the dataset to the closest medoid ∂_j.
4. Compute the mean of the dissimilarities of the observations compared to their closest medoid. This is used as a measure of the strength of the cluster.
5. Retain the subset of data for which the mean is minimal.
6. Repeat until the objective function converges.

Appendix 2

Considering a sample of data: $d = \{x_i, y_i\}$ where $x_i \in \mathbb{R}^n$ are the input data and $y_i \in \mathbb{R}$ are the output data, a short overview of these models is provided in the following:

LR: considers a linear relationship between the independent and dependent variables: $Y = a_0 + a_1x_1 + ... + a_nx_n + \varepsilon$, where α_0 is the intercept, α_n is the nth feature coefficient and ε represents the slope [16]. The aim is to find the best intercepts values of fit straight line $Y = \hat{\alpha}_0 + \hat{\alpha}_1x_1 + ... + \hat{\alpha}_nx_n$ for a continuous response variable Y. For $x = x_i$, LR becomes an optimization problem of (*) that minimizes the slope:

$$(\alpha_0, \alpha_1, ..., \alpha_n) = Arg\min \sum_{i=1}^{n} [y_i - (\hat{\alpha}_0 + \hat{\alpha}_1x_1 + ... + \hat{\alpha}_nx_n)] \quad (*)$$

BR: enables to interpret Y sought as a random variable with probability distribution $p(D \setminus Y)$ characterizing a data model D. We assume that we have an 'Evidence' or prior knowledge about Y, expressed by the prior distribution $p(Y)$. The likelihood $p(D \setminus Y)$ denotes the probability of observing D if y is true. From $D = \{d_1, ..., d_n\}$, we obtain a posterior distribution via the Bayes rule, where Y is the response, D the observations, P(D\Y) the data likelihood, $p(Y \setminus D)$ the posterior probability density and $p(D)$ the evidence. Then, the principle consists in determining the most probable Y (3).

$$Y : Arg\max Yp(Y \setminus D) = Arg\max Yp(D \setminus Y)\, p(Y) \quad (**)$$

ANN: Each of the layers is connected by a weight summation process to activate information transfer. Then, $f(\sum_{i=1}^{n} w_{ij}x_{ij})$ will be the output of a set of the signals entering the neuron i via the activation function f, where the connection (i, j) is weighted by w_{ij}. Across calculating the error in the input units, the method aims to minimize the square error.

RF: The training algorithm for random forests is based on the general bootstrap aggregation (bagging) technique. Given a training feature $X = x1, ..., xn$ with output $Y = y1, ..., yn$, bagging repeatedly (B times) selects a random sample with replacement of the training set and fits trees to these samples. For $b = 1, ..., B$, a set of samples is provided, with replacement, n training examples from X, Y; call these Xb, Yb. Then, it trains a classification or regression tree fb on Xb, Yb. After training, predictions for unseen samples x' can be made by averaging the predictions from all the individual regression trees on x': $\hat{f} = \frac{1}{B}\sum_{b=1}^{B} fb(x')$.

SVM: The basic theory of SVM for the regression algorithm is to train the sample and use the regression function: $y(x) = w^T\vartheta(x) + b$, we denote that $w \in \mathbb{R}$. The parameter $b \in \mathbb{R}$ is estimated by minimizing the regularized function [15]. It uses the kernel function in the case of a non-separable linearly data to obtain linear and smooth surfaces. To obtain a tolerate noise amount SVM model is fitted by adding the lower training error ε_i^* and ε_i (the higher): $\min_w 1/2w^Tw + c\sum_i(\varepsilon_i + \varepsilon_i^*)$; subject to $y_i - \hat{y}_i \prec \varepsilon_i + \varepsilon_i^*$, and $\varepsilon_i, \varepsilon_i^*$, where $\vartheta(x)$ is the function that maps x_i to a high dimensional feature space and c represents the regularization constant.

KNN: with the data set using labels (x_i, y_i), the predicted output y_i is find by the closest K neighbor to the data points, and then it will allocate a class to the data points x_i with the largest probability: $P(Y = j \setminus X = x_0) = \frac{1}{k}\sum_{i \in N_0} I(y_i = j)$

Where N_0 is the set of k-nearest observations and $I(y_i = j)$ is an indicator variable that evaluates to 1 if a given observation $(x_i; y_i)$ in N_0 is a member of class j, and 0 if otherwise. After estimating these probabilities, k-nearest neighbors assign the observation x_0 to the class which the previous probability is the greatest.

CART: the relationship between an outcome item y and a feature x is defined as follow: $y = f(x) = \sum_a^A c_a I\{x \in V_a\}$. This means that each instance falls into exactly one leaf node (=subset V_a). $I\{x \in V_a\}$ is the identity function that returns 1 if x is in the subset V_a and 0 otherwise. If an instance falls into a leaf node V_l, the predicted outcome is $y = c_l$, where c_l is the average of all training instances in leaf node V_l.

Figure 10: Methodological differences between used forecasting algorithms.

6. Conclusion

Urban trade in goods rises significantly in line with the economic and demographic development of cities. The essential use of road transport increases the rate of congestion. The presence of heavy-duty vehicles that are poorly filled amplifies this phenomenon. This leads to long journeys, further reducing the efficiency of delivery drivers. Logistics consolidation centres play an important role in tackling these problems. However, from

our point of view, it is complicated to arbitrarily pursue the analysis of location sites over the entire area. In order to reduce the size of the problem and to take into account the particularities of each territory, we therefore use spatial approaches that aim to divide any area into homogeneous zones. The idea is to build clusters representing the geographical distribution of transport demand.

Each zone will be assigned a weight that reflects the importance of the logistical movements generated by the delivery points it contains. These movements can be quantified using a field survey. The expected result is in the form of an urban zoning that models the area of influence of each hub, the radius size of which is controlled by two criteria:

- The relative distance between delivery points (PDL);
- The local density of demand.

Following our future perspectives, this article proposed a machine learning approach for the creation of districts in the city. This research is complemented by a set of challenging benchmarks to improve the overall effectiveness of the approach. Naturally, some of these machine learning models were not the most appropriate, but certainly we proved that the idea is feasible. By focusing on a case study, these findings motivate us to exploit its potential in real-world city data sets.

References

Austin, S. B., Melly, S. J., Sanchez, B. N., Patel, A., Buka, S., Gortmaker, S. L., (2005). "Clustering of Fast-Food Restaurants Around Schools: A Novel Application of Spatial Statistics to the Study of Food Environments ". *American Journal of Public Health*, 95(9), 1575–1581.

Baro, J., Bonin, O., Hubert, J. P., (2016). "Élaboration d'un zonage de tissus urbains: introduire de la structure dans un référentiel carroyé ". *Revue Internationale de Géomatique*, *Lavoisier*, 26(1), pp. 33-53.

Bianco, V., Manca, O., & Nardini, S. (2013). "Linear Regression Models to Forecast Electricity Consumption in Italy". *Energy Sources, Part B: Economics, Planning, and Policy*, 8(1), pp. 86–93. https://doi.org/10.1080/15567240903289549

Carlsson, J. G., & Devulapalli, R., (2013). "Dividing a Territory Among Several Facilities ". *INFORMS Journal on Computing*, 25(4), pp. 730–742.

Delaitre, L., Breuil, D., Molet, H., (2008). "Systems Science for selecting urban freight solution: Application to La Rochelle". 2008 2nd International Conference on Research Challenges in Information Science. Marrakech, Morocco, 401–408.

Ducret, R., Lemarié, B., Roset, A., (2016). "Cluster Analysis and Spatial Modeling for Urban Freight. Identifying Homogeneous Urban Zones Based on Urban Form and Logistics Characteristics ". Transportation Research Procedia, 12, pp. 301–313.

El Ouadi, J., Errousso, H., Benhadou, S., Medromi, H., & Malhene, N. (2020). "A Machine-Learning Based Approach for Zoning Urban Area in Consolidation Schemes Context". 2020 IEEE 13th International Colloquium of Logistics and Supply Chain Management (LOGISTIQUA), 1–7. https://doi.org/10.1109/LOGISTIQUA49782.2020.9353901

El Ouadi, J., Malhene, N., Benhadou, S., & Medromi, H. (2020). "Strategic zoning approach for urban areas: Towards a shared transportation system". Procedia Computer Science, pp. 170, 211–218. https://doi.org/10.1016/j.procs.2020.03.027

El Ouadi, J., Malhene, N., Benhadou, S., & Medromi, H. (2021). Shared public transport within a physical internet framework: Reviews, conceptualization and expected challenges under COVID-19 pandemic. *IATSS Research*. https://doi.org/10.1016/j.iatssr.2021.03.001

Galvão, L. C., Novaes, A. G. N., Souza de Cursi, J. E., Souza, J. C., (2006). "A multiplicatively-weighted Voronoi diagram approach to logistics districting ". *Computers & Operations Research*, 33(1), 93–114.

Huang, Y., Savelsbergh, M., Zhao, L., (2018). "Designing logistics systems for home delivery in densely populated urban areas ". *Transportation Research Part B: Methodological*, 115, pp. 95–125.

Iliopoulou, C., Kepaptsoglou, K., & Vlahogianni, E. (2019). "Metaheuristics for the transit route network design problem: A review and comparative analysis". *Public Transport*, 11(3), pp. 487–521. https://doi.org/10.1007/s12469-019-00211-2

Janjevic, M., & Ndiaye, A. B. (2014). "Development and Application of a Transferability Framework for Micro-consolidation Schemes in Urban Freight Transport". Procedia - Social and Behavioral Sciences, 125, pp. 284–296. https://doi.org/10.1016/j.sbspro.2014.01.1474

Kazemzadeh-Zow, A., Zanganeh Shahraki, S., Salvati, L., Samani, N. N., (2017). "A spatial zoning approach to calibrate and validate urban growth models ". *International Journal of Geographical Information Science*, 31(4), pp. 763–782.

Limbourg, S., Jourquin, B., (2007). "Rail-Road terminal locations: aggregation errors and best potential locations on large networks". *European Journal of Transport and Infrastructure Research*, 19, pp. 317-334.

Manganelli, B., Murgante, B., (2012). "Spatial Analysis and Statistics for Zoning of Urban Areas ". World Academy of Science, Engineering and Technology, Open Science Index 71, *International Journal of Humanities and Social Sciences*, 6(11), 2747 - 2751.

Martin, L.-C. (2019). "Machine Learning vs Traditional Forecasting Methods: An Application to South African GDP (12/2019; Working Papers)". Stellenbosch University, Department of Economics. https://ideas.repec.org/p/sza/wpaper/wpapers326.html

Novaes, A. G. N., Souza de Cursi, J. E., da Silva, A. C. L., Souza, J. C., (2009). "Solving continuous location–districting problems with Voronoi diagrams ". *Computers & Operations Research*, 36(1), 40–59.

Ros-McDonnell, L., de-la-Fuente-Aragón, M. V., Ros-McDonnell, D., & Cardós, M. (2018). "Analysis of freight distribution flows in an urban functional area". *Cities*, 79, pp. 159–168. https://doi.org/10.1016/j.cities.2018.03.005

Rulence, D., (2003). "Gestion des réseaux de points de vente: l'importance de la dimension spatiale ". Recherche et Applications en Marketing, 18(3), pp. 65–80.

Sakai, T., (2016a). "Location choice models of urban logistics facilities and the impact of zoning on their spatial distribution and efficiency". The Proceedings of the 95th Annual Meeting of the Transportation Research Board, Washington D.C.

Segura, E., Carmona-Benitez, R. B., & Lozano, A. (2014). "Dynamic Location of Distribution Centres, a Real Case Study". Transportation Research Procedia, 3, pp. 547–554. https://doi.org/10.1016/j.trpro.2014.10.010

Wang, D., Sun, J., Dong, A., Zhu, G., Liu, S., Huang, H., & Shu, D. (2019). "Prediction of core deflection in wax injection for investment casting by using SVM and BPNN".

The International Journal of Advanced Manufacturing Technology, 101(5–8), 2165–2173. https://doi.org/10.1007/s00170-018-3069-4

Wygonik, E., & Goodchild, A. V. (2018). "Urban form and last-mile goods movement: Factors affecting vehicle miles travelled and emissions". *Transportation Research Part D: Transport and Environment*, 61, pp. 217–229. https://doi.org/10.1016/j.trd.2016.09.015

Xin, J. (2016). "Bus Dwell Time Prediction Based on KNN". Procedia Engineering, vol. 137, pp. 283–288, 2016, doi: 10.1016/j.proeng.2016.01.260.

Xu, H., Ma, C., Lian, J., Xu, K., & Chaima, E. (2018). "Urban flooding risk assessment based on an integrated k-means cluster algorithm and improved entropy weight method in the region of Haikou, China". *Journal of Hydrology*, 563, pp. 975–986. https://doi.org/10.1016/j.jhydrol.2018.06.060

Yu, B., Wang, H., Shan, W., & Yao, B. (2018). "Prediction of Bus Travel Time Using Random Forests Based on Near Neighbors". *Computer-Aided Civil and Infrastructure Engineering*, 33(4), 333–350. https://doi.org/10.1111/mice.12315

Zhao, L., Zhao, Y., Hu, Q., Li, H., & Stoeter, J. (2018). "Evaluation of consolidation center cargo capacity and loctions for China railway express". *Transportation Research Part E: Logistics and Transportation Review*, 117, pp. 58–81. https://doi.org/10.1016/j.tre.2017.09.007

Acknowledgements

The authors would like to express their gratitude to the anonymous reviewers for their constructive comments, which have significantly helped improve this article.