Predictive analysis of urban waste generation for the city of Bogotá, Colombia, through the implementation of decision trees-based machine learning, support vector machines and artificial neural networks

Johanna Karina Solano Meza a,*, David Orjuela Yepesa a, Javier Rodrigo-Ilarri b, Eduardo Cassiraga b

a Department of Environmental Engineering, Santo Tomás University, Cra. 9 # 51-11, Bogotá, Colombia
b Water and Environmental Engineering Institute (IIAMA), Polytechnic University of Valencia, Camino de Vera, s/n, 46022 Valencia, Spain

ARTICLE INFO

Keywords:
Environmental science
Waste treatment
Water treatment
Green engineering
Environmental chemical engineering
Waste
Urban solid waste
Artificial intelligence
Urban solid waste management
Tree through machine learning
Support vector machines
Artificial neural network

ABSTRACT

This study presents an analysis of three models associated with artificial intelligence as tools to forecast the generation of urban solid waste in the city of Bogotá, in order to learn about this type of waste's behavior. The analysis was carried out in such a manner that different efficient alternatives are presented. In this paper, a possible decision-making strategy was explored and implemented to plan and design technologies for the stages of collection, transport and final disposal of waste in cities, while taking into account their particular characteristics. The first model used to analyze data was the decision tree which employed machine learning as a non-parametric algorithm that models data separation limitations based on the learning decision rules on the input characteristics of the model. Support vector machines were the second method implemented as a forecasting model. The primary advantage of support vector machines is their proper adjustment to data despite its variable nature or when faced with problems with a small amount of training data. Lastly, recurrent neural network models to forecast data were implemented, which yielded positive results. Their architectural design is useful in exploring temporal correlations among the same. Distribution by collection zone in the city, socio-economic stratification, population, and quantity of solid waste generated in a determined period of time were factors considered in the analysis of this forecast. The results found that support vector machines are the most appropriate model for this type of analysis.

1. Introduction

The current problem associated with increasing urban solid waste (USW) generation in large cities requires a constant search for strategies and methods that enable its proper management, while taking into account the distinct characteristics of each zone of a particular city. Municipal solid waste management represents an increasingly important environmental, economic, and social challenge for cities. Understanding behavioral patterns in the generation of household waste is a critical component for efficient collection, and to design incentives that encourage recycling and composting (Kontokosta et al., 2018). "The factors that affect the amount of generated waste are: geographical situation, seasonal fluctuation, collection system, onsite process, people's food habits, economic conditions, recovery and reuse boundaries, existing laws, and people's cultural conditions," (Roy et al., 2013). In addition to the aforementioned, a growing population, increases in industrialization and urbanization, a lack of awareness on the part of the population regarding waste generation, the public's lifestyle, temperature, annual precipitation averages, a lack of funds for sanitation, insufficient quantity of vehicles and projects in this area, and poor planning are all contributing factors responsible for poor waste management (Singh and Satija, 2016). The city of Bogotá, the capital of Colombia, is a city whose complexity and dynamism requires the implementation of technologies, policies and guidelines to properly manage urban solid waste. This is due to the characteristics of its growing population, which have diverse socio-economic levels that behave differently.

The development of a municipal solid waste management plan is a complex process. Developing an efficient plan, quantifying and forecasting solid waste generation are essential components. Forecasting waste generation cannot be done directly, as it depends on many factors. In practice, due to uncertainty and unavailability of sufficient data, modeling methods are needed to forecast USW generation (Kolekar et al.,

---

* Corresponding author.
E-mail address: johannasolano@usantotomas.edu.co (J.K. Solano Meza).

https://doi.org/10.1016/j.heliyon.2019.e02810
Received 23 July 2019; Received in revised form 17 September 2019; Accepted 7 November 2019
2405-8440/© 2019 The Author(s). Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).
Having the ability to forecast USW, with minimum errors, in different zones of a city, enables local waste management entities, as well as different government agencies to make decisions related to each management stage. This result in the proper management, use and disposal of waste while minimizing negative environmental impacts.

Artificial intelligence models perform well in prediction tasks and could be successful in establishing forecasting models for municipal solid waste. The use of machine learning algorithms can reliably predict monthly USW generation by being trained through a waste generation time series. "In the scientific community, there are publications that report the application of information and communication technologies (ICTs) and methods of operations research (OR) for solid waste management (SWM). The methods that help the decision-making process, such as multi-criteria decision analysis (MCDA) linked to ICTs, favor the development of efficient solutions for planning and management, while assisting managers in environmental issues," (Vitórin et al., 2017).

1. ANNs have demonstrated that their application to forecast the average rate of seasonal municipal solid waste generation produces accurate results, in which the difference between the predicted and observed values is not statistically significant. When compared with multiple linear regression models, ANNs demonstrate greater precision due to their non-linear nature (Azadi and Karimiashni, 2016). Moreover, decision trees are a non-parametric method used for recognizing complex patterns, particularly in classifying tasks that involve several types of patterns and a large number of attributes. Their main objective is the successive evaluation of a decision function in a manner that reduces uncertainty in identifying an unknown pattern (Goddard et al., 1995). While support vector machines are based on local data separation functions called “kernels,” "in which data is mapped into a higher dimensional space, which seeks the maximum separation between classes, as a classifier of a single class. The description given by the data from the supporting vectors is capable of forming a decision limit around the domain of the learning data with very little or no knowledge of the data outside the boundary. This boundary function, when brought back into the input space, can separate data into each different class, each one forming a cluster," (Betancourt, 2005). Support vector machines are therefore presented as a convenient tool for these types of studies in which the quantity of data is limited by different related processes. This tool can lead to better results if used with the principal component analysis technique (PCA) (Noori et al., 2009). Thus, this comparative analysis of predictive models for the generation of solid urban waste is of great importance, as it enables the determination of the most appropriate tool. This analysis supports municipal administrations by enabling them to make better decisions, effectively minimizing negative environmental impacts through their management. Furthermore, it contributes to strengthening the practical inclusion of this type of tool in this area of knowledge.

2. Materials and methods

The city of Bogotá is divided into 20 administrative districts, each of which has its own mayor and a local administrative council. The city currently has a home waste collection scheme divided into 6 collection areas (exclusive service areas – ASED), as shown in Table 1. This division enables the public cleaning service to have greater coverage. Each ASE groups together a certain number of districts according to their location and provides a selective collection scheme for recyclable waste. Solid waste generation data was provided by the Special Administrative Unit of Public Services (UAESP, 2017). Data was aggregated by collection area and its respective operating company, while taking into account the administrative division of the city into districts.

The database used to create the models is made up of data corresponding to the monthly generation of solid waste by locality for 2012–2016. This time period had the largest amount of available data corresponding to the population by locality, such as socio-economic stratification. To begin, the existing data was restructured so that it could be algorithmically processed and used. The data was organized into three CSV files by category for easier manipulation, in order to analyze and implement the predictive models. As such, the first information file was created for each collection zone, which was organized in a series of contiguous rows corresponding to each observation year, in terms of population growth and the amount of solid waste generated. Afterwards, a data structure file by locality was created, which corresponds to the same information by zone, but with an identifier for each locality included (locality ID), which was the only difference. Lastly, a file was created that contains general information of the data to connect the zone and locality files, as well as their socio-economic stratification with their corresponding labeling. This process facilitated data analysis, visualization and processing, to which, longitude and latitude information was added.

Python was used to carry out the data analysis and implement the forecasting models. Notebooks technology was employed to visualize and display the data, which codes interactive digital files that envelop both

| Zone        | Locality             |
|-------------|----------------------|
| ASE 1       | USAQUEN              |
| ASE 2       | SUBA                 |
| ASE 3       | FONTIBÓN             |
| ASE 4       | ENGATIVA             |
| ASE 5       | CHAPINERO            |
| ASE 6       | SANTA FE             |
|             | BARBIOS UNIDOS       |
|             | LA CANDELARIA        |
|             | LOS MARTIRES         |
|             | TEUSAQUILLO          |
|             | CIUDAD BOBÍA         |
|             | PUENTE ARANDA        |
|             | TUNJUELOTE           |
|             | SAN CRISTOBAL        |
|             | USME                 |
|             | ANTONIO NARINO       |
|             | RAFAEL URIBE         |
|             | BOSA                 |
|             | KENNEDY              |

Fig. 1. Waste collected annually by (ASE).

Table 1: Distribution of localities by solid waste collection zone in the city of Bogotá (UAESP, 2017), January 2018.
Fig. 2. Annual growth and reduction in waste generation and population (February).
the models’ code and description, while visualizing the results obtained. This facilitates the analysis and manipulation of the information contained therein. Information in these types of tools may be analyzed separately and correlated among the same. This comparative behavioral analysis of models led to the determination that, for this type of forecasting with a limited amount of data, support vector machines are more efficient given the type and quantity of available information.

With respect to data structuring, improvements were made to the data to achieve an automated reading of the same, and to be able to demonstrate its quality in terms of its missing elements. A characteristic of solid waste management databases is the limited amount of data, as well as their continuity issues. Estimates were made of the missing data based on observations with the same characteristics. For example, values from prior years or neighboring values in the data column were used. Once the data was organized and structured, a visual analysis tool was then developed. A file was developed in Notebook that has the reading, description and visualization of the data.

The information obtained is displayed in a dynamic manner, which facilitates its analysis and manipulation. The information can be separately analyzed by zone and locality or correlated within the same.
Initially, data exploration was performed to determine relationships and patterns of the stored figures. The best method to visualize and understand data is to convert stored data into a time series. The respective time series for each zone are shown in Fig. 1, which displays the zones that generate the most and least amount of waste. Similarly, peaks and small periodic ranges can be identified. For example, the zones that generate the most waste are 1 and 6, while 3 and 4 generate the least amount. Furthermore, there are decreases in the final months of each year.

Moreover, the decision was made to analyze the time series of the localities for each month. Fig. 2 shows the relationship between the population and waste generation over the registered years (2012–2015). One can easily see that the population of each locality has grown steadily without abrupt changes, and that the locality of Suba is generating waste at an increasing rate. For this step, Fig. 3 displays an alternative visualization in function of the year and the selected month. In this case, the solid waste generation in October 2016 is shown.

Once the work files were organized, an analysis was performed on the generation of solid waste in the city. Graphs that provided a better visualization of its behavior were used for the analysis. This was followed by determining the most optimal mathematical models, with regards to the type and amount of data, as well as its behavior for each available data series. These included the amount of solid waste generated, population, socio-economic stratification, latitude and altitude.

Taking into account the limited size of the data, as well as its temporal relationship, the decision was made to use three different models: decision trees (DT), support vector machines (SVM), and recurrent neural networks (LSTM).

In this study, the mean root square error related to the value predicted by the implemented method and the actual value was used to quantify the error associated with each prediction made by this method. This error measures local differences and provides an estimate of the quantified variable, i.e., the amount of waste.

3. Results and discussion

The modeling results using the aforementioned data is shown below. First, the behavior of the results of each model was analyzed. Then, taking into consideration this behavior, a predictive analysis of solid waste generation in the city was carried out.

Fig. 5. Performance comparison of support vector machines and models with and without a sliding window.
3.1. Forecasting with decision trees

Decision trees model data separation limitations based on learning decision rules made for the model’s input characteristics. The depth of the learning tree is determined based on the required sensibility of the model, which contains a set of decision thresholds that separate the data contained in each characteristic. For example, in this particular problem, the input characteristic is waste production per year. If one decision tree is defined from two trees, the learning algorithm will attempt to separate the waste into two subgroups and there would only be one separation value. As the determined depth increases, the separation becomes narrower and the data can be separated in a more optimal manner. However, a very large tree depth value makes separation lines on the adjustment, which limits the generality of the model for new values. This means that the model learned noisy input data and will not be able to forecast coherent values for values with a determined variance with respect to the training data. Decision trees perform recursive partitioning of the characteristics space \( \{x_i\}_{i=1}^{m} \), which in this case, corresponds to the waste percentage. Each candidate partition is defined as \( \theta = (j, r_n) \), where \( j \) is the value in terms of waste and \( r_n \) is the prediction threshold. Each tree node \( Q \) is therefore defined as the partition \( Q_{\text{left}} (\theta) \) and \( Q_{\text{right}} (\theta) \). An optimized version of the CART algorithm (classification and regression trees) was used in the implementation. Fig. 4 illustrates the use of the decision tree for ASE4, using tree depths of 3 and 6.

The main advantage of this method is the speed in calculating the representative tree and the ability to visualize and understand the cuts made by the model. However, if the data is highly variable, very scarce or non-continuous, this method may be restricted in providing suitable results if compared with other predictive analysis tools.

3.2. Forecasting with support vector machines (SVM)

As stated above, the second method implemented as a forecasting model was support vector machines. This method is based on local separation functions. “SVMs are one of the most effective mathematical approaches for both machine learning and data mining communities,” (Abbas et al., 2019). These functions interpose a separation hyperplane between a set of nearby data (local), with its separation flexibility depending on the type of function. To adjust these functions, a data set called support vectors is randomly selected on a range of data. “One of the major advantages of SVM is that it can find the pattern of nonlinear input/output data,” (Anbari et al., 2015).

Once the support vectors are selected, the aim is to maximize the distance between them, called the classification margin. A regression model was calculated in this manner using a radial basis function as a kernel. Its behavior is displayed in Fig. 5. Results obtained show that this method properly adjusts to the selected data and achieves consistent regression curves despite very limited training data.
SVMs demonstrate sound performance in real-world applications, including disease detection, activity recognition, speaker identification, digit recognition, and text categorization (Tang et al., 2019). As such, it also represents an alternative for the predictive analysis of urban waste generation. The primary advantage of the SVM model is its fit with the data despite its variable nature, or when faced with problems related to limited training data. Moreover, different kernels can be used, which can better interpret training data for improved forecasting. Thus, the results show suitable behavior in the analysis according to the variables established and the amount of data provided.

3.3. Forecasting with recurrent neural networks

As a third forecasting alternative, recurrent neural network models were implemented to forecast data. “Modern methods for forecasting include expert systems, fuzzy systems, evolutionary programming, artificial neural networks (ANN) and various combinations of these tools,” (Batinić et al., 2011). There is a large variety of neural networks with varying purposes and different architectures according to the established prediction task. “Recurrent neural networks (RNNs) are a type of neural networks with cyclic connections. This configuration makes them a more potent instrument for sequence modeling than feed-forward neural networks,” (Zia and Zahid, 2019).

The decision to use these networks on this study was based on their direct application on time series prediction. One of the main advantages of these networks (also known as long-term memory networks, LSTM) is their ability to adjust non-linear data behavior and maintain memory and forget states which take into account past time information.

In this case, neurons are called memory blocks that are connected through different layers. Each block has three gates defined as:

- Forget gate: decides which relevant information must be kept to predict new values.
- Input gate: decides which values will be updated to bring the network’s memory up to date.
- Output gate: conditionally decides the output that must be sent to the next memory block.

Forecasting the precise amount of solid waste is a difficult task given the diversity of the several parameters that affect its generation and management. This results in a high degree of fluctuation for forecasting. Therefore, applying a neural network as an intelligent system may be a sound option (Shamshiry et al., 2014). Among the main advantages of artificial neural networks is that they are able to solve non-linear problems and it is not necessary to know their mathematical details, as they only require familiarity with the working data (Ponce, 2010).

Algorithms based on deep neural networks use a descending gradient method to ascertain the weight between the neurons’ layers and

![Fig. 7. Performance comparison of an LSTM network with a window and models with and without sliding windows in pre-processing.](image)
connections, in this case, memory blocks. As they are normally used in conjunction with large volumes of data, there are new learning algorithms that aim to learn these parameters in a recursive manner on small data sets. As a result, these algorithms require the definition of new parameters, such as batch and epochs.

The batch refers to a subset of data selected to find the optimization gradient where the algorithm will converge. Epochs on the other hand, refer to the number of iterations that the descending gradient executes in the objective function to minimize the problem and find a solution. To improve the behavior of these methods, a sliding window was implemented beforehand that temporarily filters data to obtain a sequential smoothing based on local neighbors. Figs. 6, 7, and 8 display the results obtained.

Taking into account that LSTM networks were designed for temporal prediction problems, they have their own parameters that can be adapted for better prediction behavior. For instance, with scaled signals that have their own frequency periods, a time step can be defined.

As a specific example, if the signal is defined as monthly and several years have been recorded, one would expect a year to be the ideal signal period, with a behavior for each of the twelve months. Consequently, the time step could be adjusted to twelve.

As such, the results confirm that neural networks are a suitable model for this type of analysis, but they must be submitted to a robust database to obtain reliable and more accurate results. Therefore, for this case, while neural networks are a sound alternative, vectoral support machines are the best option given the characteristics of the data analyzed.

The last phase of this study consisted of the resulting forecasts made with the trained models, and the characterization of the forecasted waste by stratum and waste category for each locality or zone. Initially, for this interactive environment, the model type and the desired parameters had to be configured to start the training process. It should be noted that different parameters result in different outcomes. The option to configure the models enables an analysis and evaluation of the impact of different parameter values on the forecasting problem being addressed. The initial configuration process is described below:

- It was necessary to choose which type of data will be used for the forecast (either the zones in general or the specific localities)
- Next step was determining the size of the smoothing window. That is, to average or smooth the series of values by using a sliding window through which the parameter size can be chosen.
- Then, the model to train is chosen. In this case SVM and LSTM models were chosen for greater favorability according to the analysis performed.
When using SVM the following variables must be configured:

- Number of forecasts to make; how many predictions are to be made based on the last registered month.
- Number of training sets: number of iterations that the algorithm must use for the training has to be established.
- Lastly, the number of cross validations, or the number of subsets into which the data is divided into is established, in order to validate the training process.

If LSTM is the chosen model, the following decisions should be made:

- The layers that the recurrent neural network will have to facilitate learning finer details or to make further adjustments to the training data.
- The number of training sets described in the prior model.
- Internally, determining how many values (months) to take into account to forecast the following value. This refers to the time step parameters to be configured.

Once the models are trained, the waste generation forecasts for the following months is produced. Figs 9 and 10 show the forecasts for both the training set and the predictions made by the models for the localities of Chapinero and San Cristóbal in zones ASE 3 and ASE 6, using SVM and LSTM, respectively. It can be seen that the forecast made for the locality of Chapinero is much smoother and coherent with the patterns in the data set. Conversely, the model for the locality of San Cristóbal is over adjusted to the data, and the forecasts do not seem to follow the same trend. This is primarily due to the nature of the SVM model.

Regarding the forecasts made using the LSTM model, it has been seen that they tend to follow an average. This is mainly due to the fact that forecasts are made in a recursive manner. Predictions are made based on prior predictions. In general, these models need large amounts of data to maintain a history of more detailed patterns, and the network has to be continuously fed with real observations that have been gathered.
4. Conclusions

Three different prediction methods were implemented to forecast the generation of solid waste in the city of Bogotá. Work began by exploring the use of decision trees with different depths in a manner that could show the effectiveness of this tool in carrying out a comparative analysis. Vector support machines were also implemented to calculate regression models based on local radial functions, using points (support vectors) in a specific neighborhood as a secondary alternative given the possibility of obtaining suitable results with this tool due to the limited amount of data. Lastly, methods based on neural networks were implemented to estimate the trajectories of the points. This was the last tool used to perform the predictive analysis calculations. These methods are known as LSTM networks. The implementation of LSTM networks took into account different configurations that were temporary filtered as well as calculated annual periods of waste. As an error calculation tool, the mean root square error was used. The outcome was suitable behavior for the support vector machines and LSTM networks in the above examples. For instance, the vector support machines show an average error of 898 tons in zone 4, while SantoSanto Tomás University SantoSanto Tomás University LSTM yields an error of 1444 tons in the same zone. On the basis of this quantitative evidence, it can be concluded that SVM is the method with the best behavior in terms of the local trend of the points and assuming a high reliability in the recorded values. Likewise, this tool enables the correct calculation and analysis considering the limited amount of data which is often the case in comprehensive solid waste management processes in which records are limited or non-continuous due to the characteristics of the manner in which these types of practices are carried out in cities. These types of tools enable the calculation of trends for future values of waste that a city may generate to thus be able to design more precise and different types of strategies for the transport, collection, use and disposal of solid waste generated by a city.

Declarations

Author contribution statement

Johanna Solano, David Orjuela Yepes: Conceived and designed the experiments; Performed the experiments; Analyzed and interpreted the data; Contributed reagents, materials, analysis tools or data; Wrote the paper.

Javier Rodrigo Ilarri, Eduardo Cassiraga: Conceived and designed the experiments; Analyzed and interpreted the data; Contributed reagents, materials, analysis tools or data; Wrote the paper.

Fig. 10. Training set and prediction graphs for the LSTM model in zones ASE 4 and ASE 6.
Funding statement

This work was supported by the Santo Tomás University.

Competing interest statement

The authors declare no conflict of interest.

Additional information

No additional information is available for this paper.

Acknowledgements

Many thanks to the Final Disposal Area of the Special Administrative Unit of Public Services of Bogotá for their support in providing data to perform this research study.

References

Abbas, A.K., Al-haideri, N.A., Bashikh, A.A., June 2019. Implementing artificial neural networks and support vector machines to predict lost circulation. Egypt. J. Pet. Anbari, E., Adib, H., Iranshahi, D., 2015. Experimental investigation and development of a SVM model for hydrogenation reaction of carbon monoxide in presence of Co–Mo/Al2O3 catalyst. Chem. Eng. J. 276, 213–221.

Azadi, S., Karimiashni, A., 2016. Verifying the performance of artificial neural network and multiple linear regression in predicting the mean seasonal municipal solid waste generation rate: a case study of Fars province, Iran. Waste Manag. 48, 14–23.

Batinić, B., Vukmirović, S., Vujčić, G., Stanisavljević, N., Ubavin, D., Vukmirović, G., 2011. Using ANN model to determine future waste characteristics in order to achieve specific waste management targets - case study of Serbia. J. Sci. Ind. Res. (India) 70 (7), 513–518.

Betancourt, G.A., 2005. Support Vector Machines (SVM) Scientia et Technica Year XI, No. 27, April 2005. UTP. ISSN 0122-1701.

Goddard, J.C., Cornejo, J.M., Martínez, F.M., Martínez, A.E., Rufiner, H.L., Acevedo, R.C., 1995. Neural networks and decision trees: a hybrid approach. In: Proceedings of the International Computer Symposium Organized by the National Polytechnic Institute, pp. 1-7.

Kolekar, R.A., Hazra, T., Chakrabarty, S.N., 2016. Prediction of municipal solid waste generation models. Procedia Environ. Sci. 35, 238–244.

Kontokosta, C.E., Hong, B., Johnson, N.E., Starobin, D., 2018. Computers, Environment and Urban Systems Using Machine Learning and Small Area Estimation to Predict Building-Level Municipal Solid Waste Generation in Cities. In: Computers, Environment and Urban Systems, (November 2017, 0–1.

Koopathar, K., Hazra, T., Chakrabarty, S.N., 2016. Prediction of municipal solid waste generation with combination of support vector machine and principal component Analysis. A Case Study of Mashhad 28 (2).

Ponce, P., 2010. Artificial Intelligence with Applications for Engineering, first ed., 7. Alfaomega Editorial Group.

Roy, S., Rafieul, I., Didadul, M., Asma, U., Shobael, M., Hesbul, M., 2013. Prediction of municipal solid waste generation of khulna city using artificial neural network: a case study. Int. J. Eng Res-Online 1 (1).

Singh, D., Satija, A., 2016. Prediction of municipal solid waste generation for optimum planning and management with artificial neural network — case study : faridabad City in Haryana State (India ). Int. J. Sys. Assur. Eng. Manag. 1-7.

Special Administrative Unit of Public Services (2017) UAESP, 2017. Final disposal report of urban solid waste. Final disposal area, Tang, L., Tian, Y., Pardalos, P.M., 2019. A novel perspective on multiclass classification: regular simplex support vector machine. Inf. Sci. 480, 324–336.

Vitorino, A., Melaré, D.S., Montenegro, S., Faceli, K., Casadei, V., 2017. Technologies and decision support systems to aid solid waste management: a systematic review, 59, 567-584.

Zia, T., Zahid, U., 2019. Long short-term memory recurrent neural network architectures for Urdu acoustic modeling. Int. J. Speech Technol. 22 (1), 21–30.