Prediction performance of separate collection of packaging waste yields using Support Vector Machines

V Sousa[1*], I Meireles[2], V Oliveira[3,4], C Dias-Ferreira[3,4]

[1] CERIS, Department of Civil Engineering, Architecture and GeoResources, Tecnico Lisboa - IST, Av. Rovisco Pais, 1049-001 Lisbon, Portugal

[2] Department of Civil Engineering, University of Aveiro, Campus Universitário de Santiago, 3810-193 Aveiro, Portugal

[3] Research Centre for Natural Resources, Environment and Society (CERNAS), College of Agriculture (ESAC), Polytechnic Institute of Coimbra, Bencanta 3045-601 Coimbra, Portugal

[4] Materials and Ceramic Engineering Department, CICECO, University of Aveiro, Campus Universitário de Santiago, 3810-193 Aveiro, Portugal

[*]corresponding author. Tel.: +351 933 755 084 E-mail address: vitor.sousa@tecnico.ulisboa.pt

Abstract

This paper aims at evaluating the performance of support vector machines (SVM) technique in the prediction of separate collection yields for packaging waste. A robust regression analysis is also carried out in order to assess the influence of the dataset outliers on the model performance. The performance of SVM and robust regression models is compared with ordinary least squares non-linear regression (OLS-NL regression) models through a case study of 42 municipalities located in coastal area of the "Centro" region of Portugal. The coefficient of determination, R^2, was used to evaluate the performance of these models. The performance of SVM models (R^2=0.93) was 26% higher than OLS-NL regression model (R^2=0.73). The inclusion of the outliers of dataset does not improve the prediction performance of both OLS-NL regression (R2=0.65 against R2=0.73 without outliers) and robust regression models (R2=0.63 against R2=0.71 without outliers) indicating that the outliers impact more on the model performance than the type of regression technique used. The results demonstrate that SVM model can be a viable alternative for prediction of separate collection of packaging waste, being a valuable tool to waste management organizations in the definition of strategies and planning.

Keywords

Household packaging waste; multiple linear regression; separate collection; support vector

machines; Predictive model.

1. Introduction

Effective and efficient separate collection of waste materials such as glass, paper/cardboard, plastic /metal and bio-waste is an important concern for European Commission within the broader scope of promoting circular economy (European Comission, 2015). Indeed, recently, a new Circular Economy Package was presented by the European Commission setting new targets for 2030: i) reduction of landfilling to a maximum of 10% of the total municipal waste; ii) increase re-use and recycling of municipal waste to 65%; iii) increase recycling of packaging waste to 75% and iv) ban landfilling of separately collected waste (European Parliamentary Research Service, 2016). A better understanding of factors influencing the performance of source separated waste collection model will aid waste management organizations in planning and developing of strategies to optimize the service in order to meet with the above targets and, consequently, contribute to enable the development and implementation of circular economy.

Several methods have been proposed to model the separate collection of waste, namely descriptive statistical (Sha'Ato et al., 2007), regression analysis (Oliveira et al., 2018; Wei et al., 2013), material flow model (Tonjes and Greene, 2012), time series analysis (Xu et al., 2013) and artificial intelligence models (Abbasi et al., 2014; Kannangara et al., 2018; Noori et al., 2010). According to Abdoli et al. (2012), the artificial intelligence models have shown a superiority to conventional models in waste management topic. In addition, the high flexibility and prediction abilities of artificial intelligence models were recognized as the main advantage by Abbasi and Hanandeh (2016). On the other hand, the most commonly used multiple regression models does not provide a higher precision when inaccurate data are used in order to find information hidden in data source (Abbasi et al., 2013).

The machine learning or artificial intelligence approaches include support vector machine (SVM), adaptive neurofuzzy inference system and artificial neural network. These models have been successfully applied for municipal solid waste generation prediction (Abbasi and Hanandeh, 2016). The SVM is a novel neural network algorithm introduced by Vapnik (1995) and is based on statistical learning theory (Vapnik, 1988). This technique is powerful in solving problems characterized by small sample nonlinearity, high dimension and local minima (Noori et al., 2009). In addition, in the SVM models the empirical risk minimisation is replaced by structural risk minimisation principle (principle of the other neural networks) in order to minimise the upper bound of the generalisation error instead of minimise the misclassification error from the correct solution of the training data (Kim, 2003). Therefore, a right balance between the quality of the approximation of the given data and the complexity of the approximating function could originate

a global optimum network structure in SVM models (Noori et al., 2009). Within waste management, a few studies have used SVM models (Abbasi et al., 2014, 2013). A forecast model used to predict weekly municipal solid waste generation in Tehran city was reported by Abbasi et al. (2013). The authors concluded that SVM could predict municipal solid waste generation in a short period scale. The study of Abbasi et al. (2014) tested the application of Wavelet Transform pre-processing method to increase prediction ability of an SVM model. Data from seasonal waste generation in Tehran and Mashhad cities were used for modelling. The performance of the reported SVM models, assessed by the coefficient of determination (R^2), ranged between 0.702 to 0.761 which were considered acceptable by the authors.

In the present investigation, a different period scale and a different application of the SVM technique is tested. An SVM model will be built to predict separate collection of packaging waste yields using an annual time scale.

In a previous effort (Oliveira et al., 2018), separate collection of packaging waste yields in the coastal area of the "Centro" region of Portugal was modelled using ordinary least squares (OLS) simple and multiple linear (OLS-L regression) and non-linear regression (OLS-NL regression). The models were built starting from a set of 14 possible explanatory variables, which were reduced to 5 statistically significant. The non-linear regression model explained about 73% of variability on source separated collection yields data. The present contribution aims at trying to improve this result by developing alternative models using robust regression and SVM. The robust regression was selected because in Oliveira et al. (2018) three municipalities were removed from the sample used to develop the OLS regression model due to being statistically outliers. However, the data from those municipalities were provided by the same sources of information (waste management organisations, national statistics reports) of the remaining municipalities and, therefore, there is no real justification to remove them from the sample in addition to violating the assumption underlying the ordinary least squares regression. On the other hand, SVM were chosen in the current work to further explore the interaction between the factors, while in the previous work (Oliveira et al, 2018) the independence of the explanatory variables in the regression models developed was assumed. The possible interaction between the explanatory variables was unknown at the onset and testing all possible combinations using regression models is a cumbersome task, so SVM provides a valuable alternative since the evaluation of these interactions is intrinsic to its architecture.

2. Previous Research Framework

The identification of variables that may affect significantly the performance of separate collection of packaging waste is essential to provide adequate input data and build an accurate model for its

prediction. In Oliveira et al. (2018) a set of indicators were evaluated as independent variables for predicting the separate collection of packaging waste. A total of 14 independent variables were quantified for the municipalities, falling into two groups: socio-economic/demographic (namely, inhabitants; area degree of urbanization; purchase power per capita; purchase power index; deprivation index; population over 65 years old; number of school years attended; population over 15 years old without education or with only the first cycle of education) and variables related to waste collection service (namely, bring-banks per area; accessibility to separate collection services; relative accessibility to bring-banks; civic amenity drop-off sites per area; and inhabitants per bring-bank. The best OLS regression model achieved an R^2 of 0.73 using a combination of 5 independent variables: i) inhabitants per bring-bank, ii) relative accessibility to bring-bank, iii) degree of urbanization, iv) number of school years attended and v) area. This model implied the removal from the sample of the cases identified as statistical outliers and assumed the inexistence of interaction between the explanatory variables.

3. Methodology

The approach adopted aimed at enhancing the fitting of the model to simulate the performance of separate collection of packaging waste previously developed by Oliveira et al. (2018). For that purpose, two major options were explored: i) development of regression models that do not resort to the least squares algorithm; and ii) exploring the use of SVM in the model development.

3.1 Robust regression

One of the OLS assumptions is the inexistence of outliers, since they tend to deviate the least squares fit in their direction by receiving more weight. In the process of minimizing the residuals, the OLS estimates of the regression coefficients are distorted by this deviation (Huber, 2011). As such, in Oliveira et al. (2017) the common approach of using Cook's distance criterion to identify and remove outliers was adopted. However, there is no measurement or other type of error justifying their removal and they represent observed separate collection performance in the respective municipalities.

Comparing with the OLS regression, robust regression methods provide a less restrictive alternative regarding the outliers by attempting to dampen the influence of outlying cases. One of the approaches resorts to M-estimators and the weight of the outliers is reduced through the minimization of the sum of a function $\rho(\cdot)$ of the residuals:

$$\min_{\beta} \sum_{j=1}^{N} \rho\left(\frac{y_j - x_j'\beta}{s}\right) = \min_{\beta} \sum_{j=1}^{N} \rho(e_j)$$

were β are the regression coefficients, N the number of observations, y_j the observed values of the independent variable, y_j the observed values of the dependent variable, e_j the error and s a robust estimate of the scale of the error. This option was used herein, resorting to the Huber's method for the function of the residuals:

$$\rho(u) = \begin{cases} u^2 & if|u| < c \\ |2u|c - c^2 & if|u| \geq c \end{cases}$$

with $c = 1.345$. These equations are solved using the iteratively reweighted least squares algorithm.

The use of robust regression comprised of two analyses: i) comparison the performance of OLS-L and robust regression models considering all observations; and ii) comparison the performance of OLS-L and robust regression models excluding the outliers. The results of the OLS-NL regression model were also included for reference.

3.2 Support Vector Machines

Regression models assume, by default, that the effect of each explanatory variable on the dependent variable is independent of the value of the remaining. Consequently, possible interaction between the independent variables on the dependent variable is not accounted for unless variable transformation is used. However, there is no reason for the effect of any variable, for instance the number of inhabitants of bring-bank, on the separate waste collection to be the same independently of the other variables values, for instance the degree of urbanization.

Since the nature of the possible interaction is unknown, this is a highly demanding task using regression models. In this context, SVM provide a means to explore possible interactions between explanatory variables by constructing a hyperplane or set of hyperplanes in a high- or infinite-dimensional space, which can be used for classification, regression, or other tasks. Support vector machine is a linear machine using non-linear function u(x) and working into high dimensional feature space created by non-linear mapping of the N-dimensional input vector x into a K-dimensional feature space (K >N). The number of hidden units (K) is equal to the number of the learning data points (so called support vectors), closest to the separating hyperplane.

Developed by Cortes and Vapnik (1995), the SVM are supervised learning models derived from Vapnik and Chervonenkis' statistical learning theory (Vapnik and Chervonenkis, 1971). Originally, the SVM were developed for classifying linearly separable classes by selecting the linear classifiers minimizing the generalization error, or at least an upper bound on this error, derived from structural risk minimization. In this conception, the decision hyper-plane that characterizes the SVM leaves a maximum margin between the two classes, where the margin is defined as the sum of the distances of the hyper plane from the closest point of the two classes

(Vapnik, 1995). In other words, a good separation is achieved by the hyperplane that has the largest distance to the nearest training data points of any class (so-called functional margin), since in general the larger the margin the lower the generalization error of the classifier.

The underlying concept of the SVM was extended to non-separable classification problems using a soft margin. The soft margin allows for mislabelled points to be accounted for by including a slack variable measuring the degree of misclassification. In this case, the SVM tries to find the hyper-plane that maximizes the margin and, at the same time, minimizes a quantity proportional to the number of misclassification errors. The trade-off between margin and misclassification error is controlled by a positive constant included in the SVM objective function.

The efficient application to non-linear classification problems was made possible by using the kernel trick. Whereas the original problem may be stated in a finite dimensional space, it often happens that the sets to discriminate are not linearly separable in that space. For this reason, it was proposed that the original finite-dimensional space would be mapped into a much higher-dimensional space, presumably making the separation easier in that space. The kernel trick enables the linearization of a non-linear hyper-plane, while guarantying that the dot product with the vector normal to the hyper-plane is easily solved. Common kernels include linear, polynomial, radial basis function or hyperbolic tangent functions.

This logic was then expanded to regression problems by replacing the vector normal to separating hyperplanes in the classification problem by the regression coefficients. Also, the margin for regression errors in this case work to identify points to far from the regression line, rather than identify points in the wrong side of the class separation hyperplane. The regression SVM model also includes a parameter to balance the trade-off between the model complexity (flatness) and the degree to which deviations larger than the regression error margin are tolerated in the optimization formulation.

The best set of attributes to include in a SVM are not necessarily the same that for an OLS regression, since the assumptions underlying the later do not apply. In addition to develop an SVM model using the OLS regression model variables, a genetic algorithm was used to find a set of variables that improves the SVM prediction performance. The performance of models was done based on accuracy index: the coefficient of determination. A comparison of the performance obtained by OLS-NL and SVM models was carried out.

The data used in the regression and SVM models was obtained from 3 main sources: i) waste management companies annual reports; ii) the Portuguese water, sanitation and waste regulatory entity database; and iii) national statistical database. This data was already discussed in Oliveira et al. (2017).

4. Results and Discussion

4.1 Robust regression vs OLS

Robust regression analysis was carried out considering the same explanatory variables used in OLS regression models developed by Oliveira et al. (2018). The results obtained with robust regression, OLS-L and OLS-NL using dataset with and without outliers are presented in Figure 1 and Figure 2, respectively. The performance obtained by OLS-L and robust regression models with complete dataset are similar, with robust regression achieving slightly less overall R^2 with (less 0.11%) and without outliers (less 0.03%). However, since the highest R^2 was obtained with OLS-NL regression model with (R^2=0.646 against R^2=0.626 with the OLS-L and robust regression models) and without outliers (R^2=0.732 against R^2=0.709 and R^2=0.708 with the OLS-L and robust regression models), robust regression model does not improve the prediction performance in this case study. It can be concluded that the outliers impact more on the model performance than the type of regression technique used.

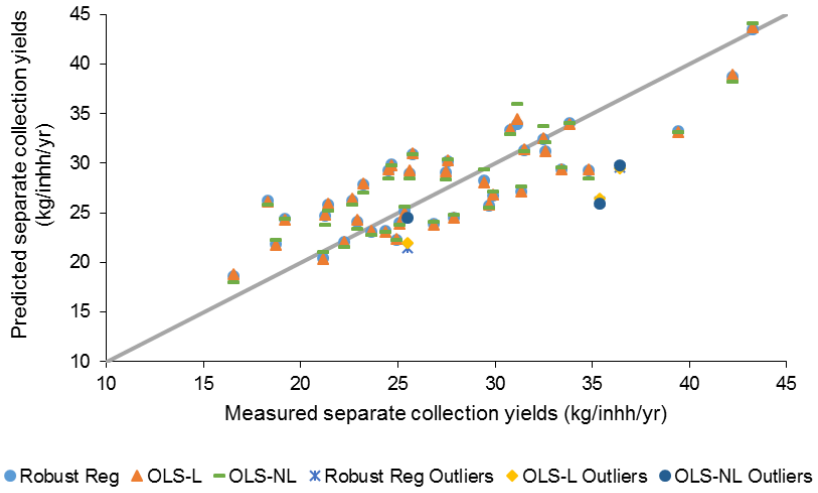

Figure 1. Comparison of predicted separate collection yields between the robust regression, OLS-L and OLS-NL models using the dataset with outliers (performance of the models predicting the outliers is presented)

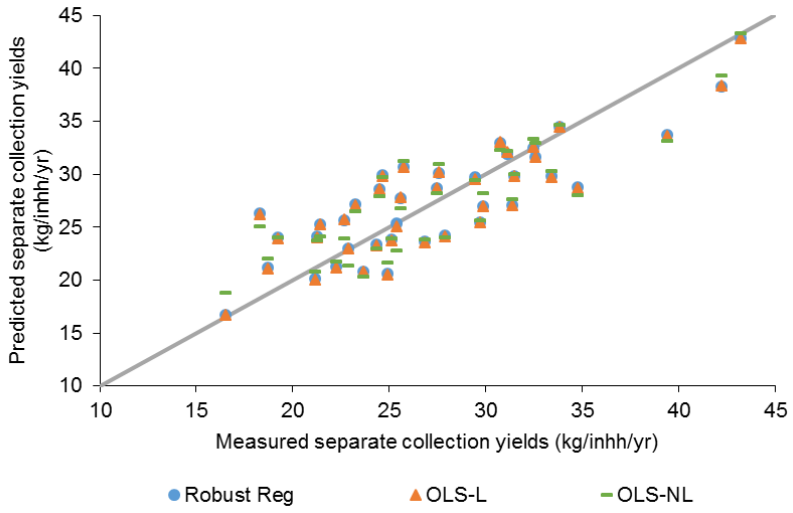

Figure 2. Comparison of predicted separate collection yields between the robust regression, OLS-L and OLS-NL models using the dataset without the outliers

4.2 Performance of SVM models in separate collection of packaging waste prediction

The SVM models were built considering dataset without the outliers. The results obtained with SVM-Reg and SVM-GA models are presented in Figure 3. In both SVM models there was a significantly increase in the explanatory power compared to OLS-NL regression model (R^2=0.732). Using the same explanatory variables of the OLS-NL regression model (SVM-Reg) or the set of variables selected using a genetic algorithm optimization (SVM-GA) yield the same performance of the SVM models (R^2=0.983). The performance of SVM model in this investigation was higher than reported by Abbasi et al. (2014, 2013) that predicted the weekly municipal solid waste generation in Tehran (R^2= 0.756 to 0.761) and Mashhad (R^2=0.702), Iran.

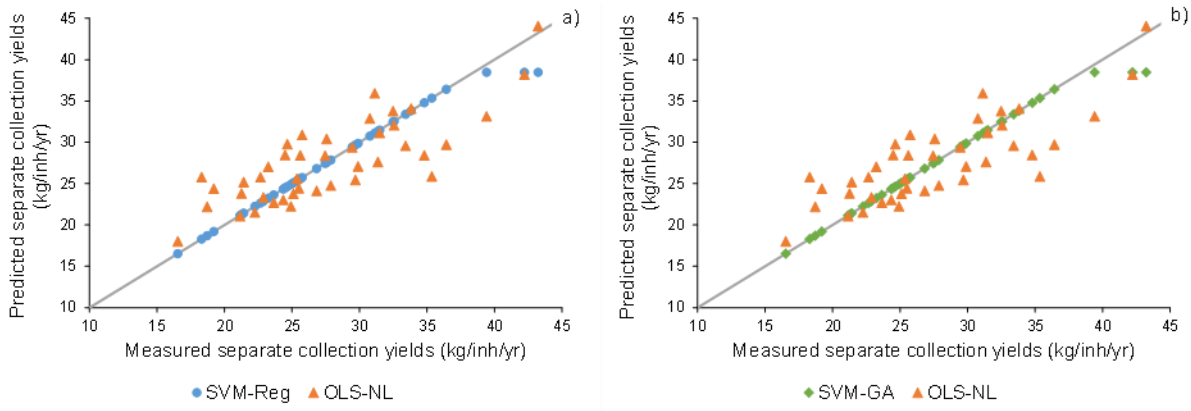

Figure 3. Comparison of predicted separate collection yields between: a) OLS-NL regression and SVM-Reg models and ii) OLS-NL regression and SVM-GA models

5. Conclusions

In this work, the performance of OLS-NL regression and SVM models for predicting separate collection of packaging waste yields is presented. The best performance was always obtained with SVM models (R^2=0.983 for both SVM-Reg and SVM-GA, considering the complete dataset) compared to OLS-NL regression models (R^2=0.73). This represents an improvement of about 26 % of the model performance when artificial intelligence tools are used. This is a strong indication that the interaction between the independent variables contribute to explain the separate collection of packaging waste yields in the coastal area of the "Centro" region of Portugal. In addition, the performance of SVM models obtained in this investigation are therefore much higher than others obtained in previous studies where municipal solid waste generation was modelled.

Exploring possible interactions between the explanatory variables easily becomes overwhelming with regression models and collinearity issues may arise. So, alternatives models, such as SVM allow embed interactions in their structure. However, this additional feature comes at a cost in terms of the model complexity, which makes a practical interpretation of the model difficult, and in terms of the generalization capability of the model to cases distinct from the ones that were used in its development. In fact, like most data-based models (also known as artificial intelligence models), SVM are prone to overfitting despite the use of techniques such as cross-validation as done in the present research. This characteristic makes the models more unstable when trying to predict values using a combination of inputs significantly different from the ones used to develop the model.

Acknowledgments

C. Dias-Ferreira, V. Oliveira and V. Sousa gratefully acknowledge FCT – Fundação para a Ciência e para a Tecnologia (SFRH/BPD/100717/2014; SFRH/BD/115312/2016; SFRH/BSAB/113784/2015) for financial support.

References

Abbasi, M., Abduli, M.A., Omidvar, B., Baghvand, A., 2014. Results Uncertainty of Support Vector Machine and Hybrid of Wavelet Transform-Support Vector Machine Models for Solid Waste Generation Forecasting. Environ. Prog. Sustain. Energy 33, 220–228. https://doi.org/10.1002/ep.11747

Abbasi, M., Abduli, M.A., Omidvar, B., Baghvand, A., 2013. Forecasting Municipal Solid waste Generation by Hybrid Support Vector Machine and Partial Least Square Model. Int.

J. Environ. Res. 7, 27–38.

Abbasi, M., Hanandeh, A. El, 2016. Forecasting municipal solid waste generation using artificial intelligence modelling approaches. Waste Manag. 56, 13–22. https://doi.org/10.1016/j.wasman.2016.05.018

Abdoli, M.A., Nezhad, M.F., Sede, R.S., Behboudian, S., 2012. Longterm Forecasting of Solid Waste Generation by the Artificial Neural Networks. Environ. Prog. Sustain. Energy 31, 628–636. https://doi.org/10.1002/ep.10591

Cortes, C., Vapnik, V., 1995. Support-vector networks. Machine Learning, 20 (3).

European Comission, 2015. Assessment of separate collection schemes in the 28 capitals of the EU (070201/ENV/2014/691401/SFRA/A2) - Final Report.

European Parliamentary Research Service, 2016. Closing the loop - new circular economy package: Briefing.

Huber, P.J., 2011. Robust Statistics. In: International Encyclopedia of Statistical science, Ed. M. Lovric, Springer, pp. 1248-1251. https://doi.org/10.1007/978-3-642-04898-2_594

Kannangara, M., Dua, R., Ahmadi, L., Bensebaa, F., 2018. Modeling and prediction of regional municipal solid waste generation and diversion in Canada using machine learning approaches. Waste Manag. 74, 3–15. https://doi.org/10.1016/j.wasman.2017.11.057

Kim, K.J., 2003. Financial time series forecasting using support vector machines. Neurocomputing 55, 307–319. https://doi.org/10.1016/S0925-2312(03)00372-2

Noori, R., Abdoli, M.A., Ghasrodashti, A.A., Ghazizade, M.J., 2009. Prediction of Municipal Solid Waste Generation with Combination of Support Vector Machine and Principal Component Analysis: A Case Study of Mashhad. Environ. Sci. Technol. 28, 249–258. https://doi.org/10.1002/ep.10317

Noori, R., Karbassi, A., Salman Sabahi, M., 2010. Evaluation of PCA and Gamma test techniques on ANN operation for weekly solid waste prediction. J. Environ. Manage. 91, 767–771. https://doi.org/10.1016/j.jenvman.2009.10.007

Oliveira, V., Sousa, V., Vaz, J.M., Dias-Ferreira, C., 2018. Model for the separate collection of packaging waste in Portuguese low-performing recycling regions. J. Environ. Manage. 216, 13-24. https://doi.org/10.1016/j.jenvman.2017.04.065

Sha'Ato, R., Aboho, S.Y., Oketunde, F.O., Eneji, I.S., Unazi, G., Agwa, S., 2007. Survey of solid waste generation and composition in a rapidly growing urban area in Central Nigeria. Waste Manag. 27, 352–358. https://doi.org/10.1016/j.wasman.2006.02.008

Tonjes, D.J., Greene, K.L., 2012. A review of national municipal solid waste generation assessments in the USA. Waste Manag. Res. 30, 758–771. https://doi.org/10.1177/0734242X12451305

Vapnik, V., 1995. Nature of statistical learning theory, Springer-Verlag. New York.

Vapnik, V., 1988. Statistical learning theory, 2nd ed, Wiley. New York.

Vapnik, V.N., Chervonenkis, A.Y., 1971. On the uniform convergence of relative frequencies of events to their probabilities. Theory Probability and Its Applications, 16 (2).

Wei, Y., Xue, Y., Yin, J., Ni, W., 2013. Prediction of Municipal Solid Waste Generation in China By Multiple Linear Regression Method1. Int. J. Comput. Appl. 35. https://doi.org/10.2316/Journal.202.2013.3.202-3898

Xu, L., Gao, P., Cui, S., Liu, C., 2013. A hybrid procedure for MSW generation forecasting at multiple time scales in Xiamen City, China. Waste Manag. 33, 1324–1331. https://doi.org/10.1016/j.wasman.2013.02.012