Forecasting Household Packaging Waste Generation: A Case Study

João A. Ferreira, Manuel C. Figueiredo, and José A. Oliveira

University of Minho, Centre Algoritmi,
Campus de Azurém, 4800-058, Guimarães, Portugal
joao.aoferreira@gmail.com,{mcf,zan}@dps.uminho.pt

Abstract. Nowadays, house packaging waste (HPW) materials acquired a great deal of importance, due to environmental and economic reasons, and therefore waste collection companies place thousands of collection points (ecopontos) for people to deposit their HPW.
In order to optimize HPW collection process, accurate forecasts of the waste generation rates are needed.
Our objective is to develop forecasting models to predict the number of collections per year required for each ecoponto by evaluating the relevance of ten proposed explanatory factors for HPW generation.
We developed models based on two approaches: multiple linear regression and artificial neural networks (ANN).The results obtained show that the best ANN model, which achieved an R^2 of 0.672 and MAD of 9.1, slightly outperforms the best regression model (R^2 of 0.636, MAD of 10.44).
The most important factors to estimate HPW generation rates are related to ecoponto characteristics and to the population and economic activities around each ecoponto location.

Keywords: Forecasting, Municipal Solid Waste Generation, House Packaging Waste, Waste Collection, Recycling, Multiple Linear Regression, Artificial Neural Network

1 Introduction

Over the last decades, the recycling of waste materials became a very important subject for society, as the environment benefits greatly from any advances made in direction of a cleaner future. In fact, as the process of recycling became really important, it also turned into an interesting resource management problem, in particular when referring to collection of waste for recycling, which involves teams of workers and vehicles. Therefore, the collection process is crucial in order for recycling to go on, and so, fleet management may frequently deal with various issues. One main problem lies in finding optimal collection routes, where a set of collection points is targeted, and each point is given a priority level. This problem can be described as a Vehicle Routing Problem (VRP), yet more flexibility is needed when it comes to choose only part of the collection points to be visited, instead of the whole set. Thus, a more fitting description of the

selective waste collection process may be the Team Orienteering Problem (TOP). In this context, the TOP can be described as the problem of designing optimized collection routes to be assigned to a fleet of vehicles that perform the collection of different types of waste stored along a network of collection points. Each one of these collection points contains a certain amount of waste that directly quantifies the respective priority level. The collection routes have maximum durations or distances, and consequently, the selection of collection points to be visited by the vehicles is made by balancing their priorities and their contributions for the route duration or distance. The objective is to maximize the total amount of waste collected by all routes while respecting the time or distance constraints.

Aside from the routing problem, there are other issues related to the process of waste collection for recycling, especially when dealing with real scenarios and the activity of real waste collection companies. One of these issues is the prediction of waste material quantities generated over time at each collection point in a given collection network, which enables the determining of a waste generation rate (WGR) for each collection point. Determining the WGR values is usually helpful during the designing phase of the collection network. Collection points are assigned to certain places based on probable WGR values that were previously assessed (predicted). Based on forecasts for waste generated along the network, the principles of the TOP can be employed to design the collection routes, as each collection point is assigned with a certain priority level according to their WGR, which translates the need for each collection point to be emptied.

In Portugal, household packaging waste (HPW) is separated by citizens at the local recycling site, named ecoponto ("ecological point"). Here, the waste is divided into three main containers, typically identified with different colours to help people to separate the waste: 1) glass (green) 2) paper/cardboard (blue), and 3) plastic/metal (yellow). These ecoponto sites are provided by the municipalities for only household waste to be recycled in these containers. Given the goals Portugal has to fulfil for the recycling and recovery of HPW, there is a permanent need for increased efficiency in waste collection.

The work presented in this paper is integrated in a R&D project named Genetic Algorithm for Team Orienteering Problem (GATOP), which is financed by the Portuguese Foundation for Science and Technology (FCT). The major goal of GATOP is the development of a more complete and efficient solutions for several real-life multi-level Vehicle Routing Problems, with emphasis on the waste collection management. Within the project scope, one important task is the development of forecasting models to predict the quantities of waste generated at collection points. In this work we focused on developing models to predict the generation of recyclable waste along a network of ecopontos. We aimed to achieve the best approximation possible of the predicted values to the real values by using regression models. We also explored other forecasting methods based on Artificial Neural Network models.

This paper is structured in 6 sections. Section 2 presents the real problem. In section 3, several forecasting methods and models found in the literature that are applied in the context of waste management are discussed. The fol-

lowed methodologies and the developed models are presented and analysed in the fourth section. Computational experiments are described in the fifth section and the results are discussed. Finally, on section 6, the main conclusions of this study are presented.

2 The Real Problem

In this paper we intend to solve a real-world problem faced by Braval, an inter-municipal waste management company in the Cávado sub region of northern Portugal. Braval takes action across six municipalities: Braga, Vieira do Minho, Vila Verde, Póvoa do Lanhoso, Amares and Terras de Bouro. Braval currently operates a network of more than 1,200 ecopontos where residents can start the recycling process of their HPW. These sites are located across the municipalities in a variety of easily accessible areas. Within the six municipalities where Braval operates, there is a mix of urban and rural areas, which prompts the demand of different strategies for waste management.

Braval's fleet does not visit all the ecoponto sites every workday, and so it is necessary to select a subset of ecopontos to visit each time route planning is done. Furthermore, given a planning horizon of, for example, a week, or a month, Braval must decide which ecoponto sites must be visited, which ecoponto sites can be visited, and which sites can be skipped during the collection routes, and then design effective routes to perform the selective collection of HPW. The priority level of an ecoponto to be visited is highly related to the amounts of waste it holds during the route planning phase, as well as their own WGR. Taking into account its priority level, an ecoponto is either selected or not to be visited during the established planning horizon.

In order to help Braval performing better route planning, reliable predictions of the amount of waste generated at each ecoponto are necessary. There are several forecasting methods and models presented in the literature that feature real-world waste management problems, and good results were achieved with those strategies on those situations. Before deciding on forecasting methodologies, we should keep in mind the kind of information resources put at our disposal by Braval. These informations consist of time series with waste collection data, more specifically records of previous waste collections performed during a period of one year, where a count was kept of how much times each ecoponto was visited and emptied during that time interval. Therefore, our main goals with this study are: 1) Determining significant factors of HPW generation as well as their relevance once applied in forecasting models; 2) Predict the number of times each ecoponto should be visited during a certain period of time, that being a week, a month, a trimester or a year. These predictions are highly related to the WGR factor, and once its value is determined for each ecoponto, all of them can be categorized into different groups based on WGR value and overall collection priority. Once the categorization is done, the obtained information can be used by the vehicle routing optimization models previously developed for the GATOP project.

3 Literature Review

The subject addressed by this study is predicting or explaining the generation of HPW for recycling purposes, but in fact, this subject is related to a larger research field which is forecasting the generation of municipal solid waste (MSW). Forecasting is a necessity for the development of waste management infrastructures. It is also important for the improvement and optimization of logistics associated to waste management. Therefore, reliable data and precise forecasts are needed in order to avoid cases of insufficient or excessive waste disposal and high or low usage of infrastructures (transportation, processing, incineration or landfilling).

A review of previously published approaches [2] revealed a great amount of methods applied to forecast MSW generation. The methods referred in the review range from purely application-oriented models to very sophisticated tools, and all of them can be identified in seven different categories: group comparison, correlation analysis, multiple regression analysis, single regression analysis, input–output analysis, time series analysis and system dynamics.

In their review, Beigl et al. [2] concluded that MSW generation is best predicted by time series analysis (when assessment of seasonal impacts is necessary), alongside correlation and regression methods. In respect to the waste generation contributing factors, Wang and Nie [12] stated that a rapid growth of the urban population and gross domestic product (GDP) were the most important ones. In an attempt to forecast MSW generation based on more factors besides the previously mentioned ones, linear regression models have been employed since the 1950's. Grossman et al. [5] enhanced forecasting methods by including in the linear regression model other factors such as: increase of population, income level and housing type. Later studies pointed out that waste generation can be related to predicted production level and consumption [3, 7], and also to private consumption [4]. More detailed analysis showed the growth of the urban population to have a greater impact than GDP on the total amount of MSW produced. Also, with factors like the increasing income and the quality of life, MSW seems to change more in composition rather than increasing in total amount generated. Other factors that may influence the generation and composition of waste are the average living standards or the average people's income, climate, living habits, level of education, religious and cultural beliefs, and social and public attitudes [1, 6].

Usually, time series forecasting models may be employed to predict MSW generation when there is access to significant amount of historical data. This method does not rely on the estimation of the social and economic factors, which can be a not so accurate procedure. Based on the comparison of several analysed forecasting methods, Beigl et al. [2] imply that a forecasting tool based on the relationship between social-economic conditions and the amount of waste generated was more suitable than a single time series analyses. In most cases, the application of modelling methods such as correlation, regression analyses, and group comparisons, seems to be the better option when the goal is to test the relationship between the level of affluence and the generation of total MSW

or a material-related fraction, and to identify significant effects of waste management activities on recycling quotas. The application of time series analyses and input–output analyses is advantageous when there is a need for special information (i.e., assessment of seasonal effects for short-term forecasts). Sorting analyses are indispensable, if impacts on the quantity of separately collected waste streams (i.e., of recyclables) are to be quantified.

After reviewing several studies from other authors within the subject of MSW or HPW generation forecasting, it was clear that all of them focused on analysing at a different level that did not match our intent, which was predicting waste generation for each collection point in a collection network, so the emphasis is on the operational level of waste collection. Our study surely pursues a different depth or degree of analysis and a more problem-specific approach, but some studies found in the literature will certainly be helpful in the process of solving the forecasting problem we presented in the previous chapter.

4 Methodology

The data we accessed consists of records with a registry of all waste collections performed by Braval in all six municipalities they operate. In more detail, these records show how much times each ecoponto in Braval's network was emptied each month during the year of 2013. Our aim is to predict, with the smallest error possible, the number of times each ecoponto needs to be emptied each year. Therefore, this number of collections per year (and per ecoponto), hereafter referred as CPY, is the dependent variable to be considered in the forecasting models to be developed. In the next subsections, a brief classification of the developed models is given, followed by a more detailed description of each one.

4.1 Model Classification

The forecasting models yet to be presented in this study can be categorized on different aspects according to "(. . .) four characteristic classification criteria: regional scale, type of modelled waste streams, type of independent variables and modelling method." as stated by [2]. In terms of regional scale, our modelling approach is certainly between household scale and settlement area scale. In respect to the type of modelled waste streams, and following the concepts referred by [2], the waste streams to be modelled are collection streams. More specifically, the aim is to model source separated waste streams related to recyclable materials such as paper/cardboard, plastics/metals and glass.

Regarding the data sources for the dependent variable, CPY, these are solely based on waste collection statistics with information extracted from Braval's reports. As for the independent variables, which are factors for the prediction of waste generation, we intend to rely only on easily accessible information that is manly included in waste management related infrastructure data sources, as well as on simple socio-economic and demographic data. Other information may be collected using some Geographic Information System (GIS). The factors we

believe to be of relevance are listed in table 1. We expect these factors might help explain the behaviour of the dependent variable. The factors will be included in the two modelling methods presented in the following subsections: Multiple Regression and Artificial Neural Networks.

Table 1. List of considered contributing factors for HPW generation.

Factor	Description	Acronym
1	**N**umber of ***E**copontos* per civil parish	NE
2	**P**opulation **D**ensity (residents per square-kilometre) in the civil parish	PD
3	**N**umber of **R**esidents per ***E**coponto*	NRE
4	***E**coponto* **D**ensity (number of ecopontos per square-kilometre)	ED
5	**E**coponto **T**ype (containers/bins can be at street level (Type 1) or underground (Type 2)	ET
6	**E**coponto **P**osition (containers/bins placed within an enclosed area (i.e. a school) or placed in open area	EP
7	**E**coponto **C**apacity (capacity of the containers/bins)	EC
8	**N**umber of **E**copontos within a **300** metres radius around considered ecoponto	NE 300
9	**D**emographic **F**actor (household density around each ecoponto in a 300 metres radius around it)	DF
10	**S**ocio-**E**conomical **F**actor (based on the number of schools, local management infrastructures, quantity of commercial activities, local attractions and relevant monuments, tourism and lodging infrastructures, leisure and sports infrastructures, restaurants, cafés and bars)	SEF

4.2 Multiple Regression Model

Regression models are widely used for predicting purposes, and have proven to be a very efficient method in many studies. The first modelling method we decided to explore was a linear regression model. Since we selected several independent variables (or factors) to explain the outcome of the dependent variable, we shall employ a multiple regression model (MRM).

In order to design the model, we started by assembling all available data in order to obtain values for all previously mentioned factors (table 1). As stated earlier, waste collection data and records were made available by Braval. Simple demographics were also found on Braval's data sources. Most of socio-economic information was obtained using GIS software tools such as Google Earth and Google Maps. The point was to use easily accessible informations to apply in forecasting models.

4.3 Artificial Neural Network Model

Even though our first choice of a modelling method relied on linear regression models, we also intended to explore other forecasting methods. We opted to develop a method based on an Artificial Neural Network (ANN), which is a machine learning technique that includes algorithms able to create some kind

of artificial intelligence with learning capabilities, and has been widely applied in forecasting models and with great success and acceptance by the research community. ANNs are networks of artificial nodes called artificial neurons, and their functioning is inspired on how the human brain works. ANNs can be used to model non-linear relationships between input and output data, and also to find patterns within extensive data.

For our ANN models, we opted to use the Multilayer Perceptron structure with only one hidden layer at first, since it is the most applied structure in similar studies [8, 9, 10, 13], and also because the complexity of our problem might not require the use of an extra hidden layer. Concerning the kind of ANN topology, a feed-forward network was chosen. The employed training methods were Back-Propagation and Levenberg-Marquardt.

5 Experiments and Results

An experimentation phase took place to explore two methodologies: Regressions models and Artificial Neural Network models. Our intent was two develop the best performing models for each adopted methodology, and thus compare the results achieved, so that we can determine the overall best model based on error performance, stability and robustness. Since we also want to determine the most contributing factors for the WGR, we tested several combinations between the factors presented in table 1. Therefore, we obtained several different models that allowed the assessment of each factor's relevance.

5.1 Data Processing

Before engaging on experiments, data processing needed to be done. We used a list of all ecopontos placed in two municipalities, Amares and Vila Verde, as well as specific characteristics of those ecopontos, such as their GPS coordinates, containers type and capacity, placement date, civil parish to where each ecoponto belongs, among others. Then, we combined those informations with the available collection statistics which referred to the number of performed collections for each recyclable waste type (paper/cardboard, plastic/metal and glass) during the year of 2013 (CPY). For this study we opted to consider only one waste stream, which was paper/cardboard, or hereafter simply referred to as cardboard stream. A distribution of the variable CPY for cardboard stream is represented in figure 1, with CPY ranging between 6 and 130, and the ecopontos are sorted in ten categories. For demographics, we used census informations to determine population density values for each civil parish. We also took record of how many ecopontos were placed in each parish.

Data treatment for socio-economic factors involved the use of GIS software such as Google Earth and Google Maps. These tools also supplied some extra demographic information referring to each ecoponto individually, by analysing their surroundings in a 300 metres radius. We used this specific length by indication of the Braval's Manager, since people tend to not travel longer distances

in order to place their HPW in ecopontos. So, considering a 300 metres radius area of effect for each ecoponto, and with assistance from the GIS software, we where able to count the number of schools, local management infrastructures, commercial activities and establishments, local attractions and relevant monuments, tourism and lodging infrastructures, leisure and sports infrastructures, restaurants, cafés and bars. These informations translated to a specific measurement that we called Socio-Economic Factor (or SEF), with values ranging on a scale from 1 to 3. The SEF values were calculated based of six sub-factors (table 2), and following equations 1 and 2.

Fig. 1. Distribution of CPY for cardboard stream in 2013.

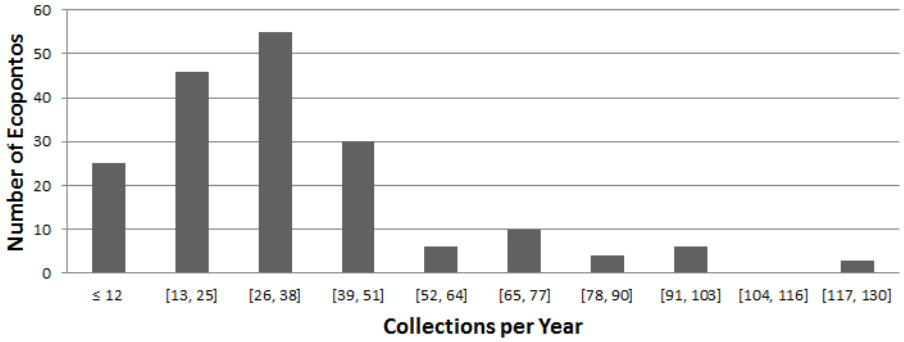


Table 2. Socio-Economic sub-factors used to determine the SEF value.

Factor	Description
SEF_1	Number of schools and education-related infrastructures
SEF_2	Number of local government infrastructures
SEF_3	Number of restaurants, cafés, bars, discos, bakery and cake houses
SEF_4	Number of shops, supermarkets, other trading-goods facilities
SEF_5	Number of hotels and other lodging facilities
SEF_6	Number of local attractions, monuments, parks, leisure and sports infrastructures

$$x = SEF_1 + SEF_2 + SEF_3 + SEF_4 + SEF_5 + SEF_6 \tag{1}$$

$$SEF = \begin{cases} 1, & \text{if } x \leq 2 \\ 2, & \text{if } 2 > x \leq 4 \\ 3, & \text{if } x > 4 \end{cases} \tag{2}$$

During the socio-economic analysis, we kept track of how many other ecopontos were placed within each ecoponto area of effect, which directly translates

to a certain competition factor between ecopontos. We assumed that a higher competition level could mean a decrease in the affluence to each ecoponto, since there are more disposal options, and the nearest option tends to be the preferred one. This situation lead to an important factor related to demographics, and we named it Demographic Factor (DF). To measure the DF for each ecoponto, its area of effect was rated on a scale from 1 to 3 (low, medium and high), in terms of households concentration. Since the values for DF depend on human-eye perception and direct observation, this factor might be susceptible to the personal arbitration of the observer. Nonetheless, an observation rule was set, so that the DF level corresponds to value 1 when up to a third of the area of effect is filled with households; the value 2 is when then household concentration ranges from one third to about two thirds of that area; finally, the DF value 3 is given when more than two thirds of the effect area has households. The DF factor is not directly related to population density, because in this case, only the populated area around the ecopontos is considered. In addition, the considered households on the map, in a top-down perspective, can either be a sole family house or a whole building with many apartments.

Not all ecopontos from Amares and Vila Verde municipalities were used in the forecasting models. Only ecopontos with a minimum amount of 6 yearly collections were considered, resulting in a total of 185 eligible ecopontos. Although some ecopontos were removed, the NE value was not updated, since these ecopontos still existed in their respective civil parish. In the following subsections, a detailed description of the experiments done with multiple regression models and ANN models are presented. Other assumptions and hypotheses are exposed and explained, and experiment results with various models are compared. While testing the models, we focused on predicting the CPY for each ecoponto considering only one waste stream: cardboard.

5.2 Experiments with Multiple Regression Model

The experiments based on a multiple regression model, hereafter MRM, were performed using the software ForecastPro. The previously treated data was transferred to the software after the ecopontos, the lines in the table, were scrambled in such a way that a balanced distribution of values was achieved, which means the first half of ecopontos was equivalent to the second half in terms of average values for waste collections, population density (PD), ecopontos in civil parish (NE), DF and SEF. For the forecasting experiments, the MRM was constructed with all ecopontos except the least 50 ones on the list, and so, the CPY values were predicted for those remaining ecopontos. Various combinations of factors resulted in different models. At first, each factor was assessed independently, resulting in ten simple regression models (just one independent variable). Later, the factors that seemed to have greater coefficient of determination (R^2) were selected and combined aiming to produce a stronger model. We also employed an inverse process of model construction, by starting with all the factors combined, and then, taking out the ones with the lowest significance to the model, from one combination to the next one. We grouped all the developed models under the

name MR1. In table 3, the most important results achieved in this first phase of experiments are presented. The considered performance measurements were R^2 and Mean Absolute Error (MAD)

Table 3. Forecast results with MR1.

MODEL	FACTORS	R^2	MAD
MR1-1	DF	0,290	15,26
MR1-2	SEF	0,219	16,25
MR1-3	DF, SEF	0,403	14,3
MR1-4	DF, EC	0,318	15,02
MR1-5	DF, ED	0,325	14,55
MR1-6	DF, EP	0,290	15,28
MR1-7	DF, ET	0,397	14,27
MR1-8	DF, NE	0,362	13,8
MR1-9	DF, NE300	0,321	14,66
MR1-10	DF, NRE	0,293	15,25
MR1-11	DF, PD	0,332	14,35
MR1-12	DF, SEF, ED	0,425	14,06
MR1-13	DF, SEF, ET	0,449	13,75
MR1-14	DF, SEF, NE	0,438	13,79
MR1-15	DF, SEF, NE300	0,417	14,22
MR1-16	DF, SEF, NRE	0,409	14,24
MR1-17	DF, SEF, PD	0,430	14
MR1-18	DF, SEF, ET, NE, PD	0,488	13,26
MR1-19	DF, SEF, ET, NE	0,487	13,23
MR1-20	DF, SEF3	0,331	14,77
MR1-21	DF, SEF4	0,367	15,2
MR1-22	DF, SEF6	0,302	15,04
MR1-23	DF, ET, NE	0,460	13,39
MR1-24	DF, ET, SEF3	0,420	14
MR1-25	DF, ET, SEF4	0,584	12,14

After analysing the results, it seemed clear that some factors stood out more than others, namely DF, SEF, NE and ET. The combination of these factors resulted in a model that achieved an R^2 value of 0.487 (model MR1-19). Regarding SEF, it seemed unclear if the way it was calculated could be masking the influence of its sub-factors (SEF1, SEF2, etc.), and so we tested each individual contribution to the model. Once tested, this hypothesis revealed a better performing model which combined DF, ET and SEF_4, with an R^2 value of 0.584 (MR1-25) and slightly lower errors, but with SEF_4 having low significance level. This result hinted us into modifying the SEF formula (equation 2). We decided to analyse the data once again, looking into the table listing all the ecopontos and their values. We paid particular attention to the SEF values and came to a conclusion: if an ecoponto had all SEF sub-factors equal to zero, it was still being awarded a SEF level equal to one, therefore, the SEF formula was modified so that when that situation happens, the SEF value would be set to zero. This modification lead to a new set of experiments but it did not translated into a superior performance of the models, considering the most relevant factors and

combinations, especially when including the SEF values. Also, the significance level of SEF in a single-variable MRM dropped. We concluded that in the MR-1 group, the socio-economic factor was not as important to the model as we thought in the beginning of the experimental phase.

Although SEF showed to have less importance compared to other factors, there were other ones that seemed to have even lower effect on the outcome of the MRM predicting performance. One of them is the ecoponto positioning (EP), and it caught our attention after another data review and analysis. Although the EP values turned to be irrelevant, the fact that an ecoponto is positioned within an enclosed area (i.e. a restaurant) or inside a facility (i.e. a school) tends to shadow the influence of other factors, since that ecoponto is mainly accessible to the users of that particular area or facility. A new hypothesis consisted on evaluating the performance of the MRM after removing the ecopontos with EP value of 1 (placed inwards), and so, a new series of tests took place, but this time only considering the remaining 185 ecopontos, and keeping the least 50 in the list for prediction of CPY. The performed tests generated a new group of forecasting models named MR2, and the most relevant results achieved are presented in table 4. This last hypothesis turned out to have a positive impact

Table 4. Forecast results with MR2.

MODEL	**FACTORS**	R^2	**MAD**
MR2-1	DF	0,526	16,62
MR2-2	DF, ED	0,535	16,53
MR2-3	DF, ET	0,571	15,88
MR2-4	DF, PD	0,534	16,55
MR2-5	DF, SEF	0,536	16,51
MR2-6	DF, EC, ED, ET,NE, NE300, NRE, PD, SEF	0,643	14,89
MR2-7	DF, EC, ED, ET, NE300, NRE, PD, SEF	0,642	14,84
MR2-8	DF, EC, ED, ET, NE300, NRE, PD	0,638	14,86
MR2-9	DF, EC, ED, ET, NRE, PD, SEF	0,636	14,92
MR2-10	DF, EC, ED, ET, NRE, PD	0,636	10,44
MR2-11	DF, ET, SEF4	0,584	15,69

on the predicting performance, and revealed other factors once considered less relevant in previous experiments, to have great influence when combined all together with DF and ET. Also, the MR2-3 model showed that with only these two factors, a strong performance was achieved in terms of R^2, with a value of 0.571, and in terms of error analysis, with lesser average error and mean absolute deviation once compared to the best performing models in the MR1 group (MR1-19 and MR1-25). This was exactly one of our goals for this study, since the aim was to develop a simple forecasting model using as few factors as possible without compromising too much the predicting performance when compared to models with several factors. Nonetheless, the overall best performing regression model is MR2-10, with a R^2 value of 0.636 and MAD equal to 10.44. All the variables used for this model presented a high level of significance, and so we can conclude the best factors to use when predicting CPY using regression methods are: DF,

EC, ED, ET, NRE and PD. The MR2-10 model is represented in equation 4. The parameter estimates for this model are presented in table 5.

Table 5. Parameter estimates for the best performing regression model - MR-10

Factor	Coefficient	Std. Error	t-Statistic	Significance
DF	18.208770	1.889130	9.638708	1.000000
EC	-0.006169	0.001458	-4.232523	0.999977
ED	8.313962	2.404589	3.457540	0.999455
ET	21.141376	3.716969	5.687800	1.000000
NIE	0.034961	0.011332	3.085258	0.997966
PD	-0.026852	0.009635	-2.786869	0.994678

$$\mathrm{CPY}_i = 18.209 \cdot \mathrm{DF} - 0.006 \cdot \mathrm{EC} + 8.314 \cdot \mathrm{ED} + 21.141 \cdot \mathrm{ET} + 0.035 \cdot \mathrm{NRE} - 0.027 \cdot \mathrm{PD} \quad (3)$$

5.3 Experiments with Artificial Neural Network Model

The development and testing of ANN models was achieved using a software tool called Neuro Solutions 6. This software offers several options in terms of ANN model construction and structuring, as well as a wide variety of training methods. For the experiments, we chose the main model characteristics based on previous studies on forecasting MSW using ANN methods [8, 9, 10, 13]. Our ANN models have the following base characteristics:

- Network Type: MLP
- Network Structure: 3 layers (input, output and one hidden layer)
- Network Topology: Feed-forward
- Training samples: 125
- Cross-validation samples: 10
- Testing samples (forecast): 50

There are several parameters that need to be tuned such as the number of neurons or processing units per layer in the ANN. Other parameters are related to the training process, and there are different options in the software for the training algorithm (learning rule) and activation functions for the neurons in the hidden and output layers. For ANN training, we opted to experiment two algorithms: Levenberg-Marquardt (L-M) and Back-propagation (BP). As for the activation functions, we opted to explore the tangent hyperbolic and sigmoid functions. The number of epochs, or training periods was set to 1000, which determines the stopping criterium for the training process. We also tested a value of 2000 to see if with a longer training could improve ANN performance. There are many possible combinations for these parameters. The most interesting ANN models and their respective results are showed in table 5. Although we also tested ANN models using the BP algorithm, the results performed obtained were worse than the models using M-L, and so we did not include them in table 6.

An analysis over the presented results showed interesting outcomes. Considering the performance indicators R^2 and MAD, the overall best model is ANN-7, yet similar results were achieved by other models such as ANN-2, ANN-3 and ANN-9. Evaluating other error measurements such as maximum absolute error, the ANN-11 model has the lowest score. So, to determine the overall best ANN model we opted to select ANN-7 by giving priority to the MAD value. The factors used for this model were only four: DF, ET, NE and SEF. We believe to have reached our goals with this ANN model, with simplicity in terms of factors used, and with a promising performance.

Table 6. Forecast results with ANN models

MODEL	ANN Config.	Activ. Function	Epochs	FACTORS	R^2	MAD	Max. Error
ANN-1	2-4-1	Tanh	1000	DF,SEF	0.624	9.80	60.34
ANN-2	2-8-1	Tanh	1000	DF,SEF	0.669	9.59	50.88
ANN-3	2-4-1	Sigmoid	1000	DF,SEF	0.672	9.55	55.28
ANN-4	2-8-1	Sigmoid	1000	DF,SEF	0.649	9.74	63.78
ANN-5	4-4-1	Tanh	1000	DF,ET,NE,SEF	0.577	11.29	50.56
ANN-6	4-4-1	Tanh	2000	DF,ET,NE,SEF	0.467	12.57	49.21
ANN-7	4-4-1	Sigmoid	1000	DF,ET,NE,SEF	0.672	9.14	64.46
ANN-8	4-4-1	Sigmoid	2000	DF,ET,NE,SEF	0.547	11.78	47.71
ANN-9	4-8-1	Sigmoid	2000	DF,ET,NE,SEF	0.682	9.19	65.06
ANN-10	6-12-1	Tanh	1000	DF,EC,ED,ET,NRE,PD	0.588	10.68	47.68
ANN-11	6-12-1	Tanh	2000	DF,EC,ED,ET,NRE,PD	0.662	11.07	42.04

5.4 Multiple Regression and ANN model comparison

The experiments conducted with the two chosen methodologies allow the comparison of regression and ANN models at predicting HPW generation, or more specifically at predicting the number of waste collections needed for each collection point in a certain network. The best regression model we developed, MR2-10, performed quite well. On the other hand, the best ANN model, ANN-7, achieved better performance than MR2-10 by obtaining higher R^2 lower MAD values. In terms of maximum absolute error in the testing sample, MR2-10 got 47.83, which is better than ANN-7 with 64.46. In figures 2 and 3, the distribution of error values within the sample (135 ecopontos) used to designed both the regression model and ANN model. In figure 4, a graphical representation of the predicted values with both ANN-7 and MR2-10 models is presented.

We considered ANN-7 the best performing model, but in ANN-3, where only two factors were applied (DF and SEF), the superior capacity of ANNs to understand relationships between variables is very clear ($R^2 = 0.672$), once compared to regression models with the same factors (i.e. MR2-5 with R^2 of 0.536). Although ANN models can achieve better performance than linear regression models, the modelling set-up was more difficult to manage, since there are many parameters involved that can greatly influence the outcome and overall success of the prediction process. Designing a neural network, deciding on the training algorithms and correctly tuning the parameters can be time consuming.

Fig. 2. Forecast results with the best ANN and regression models.

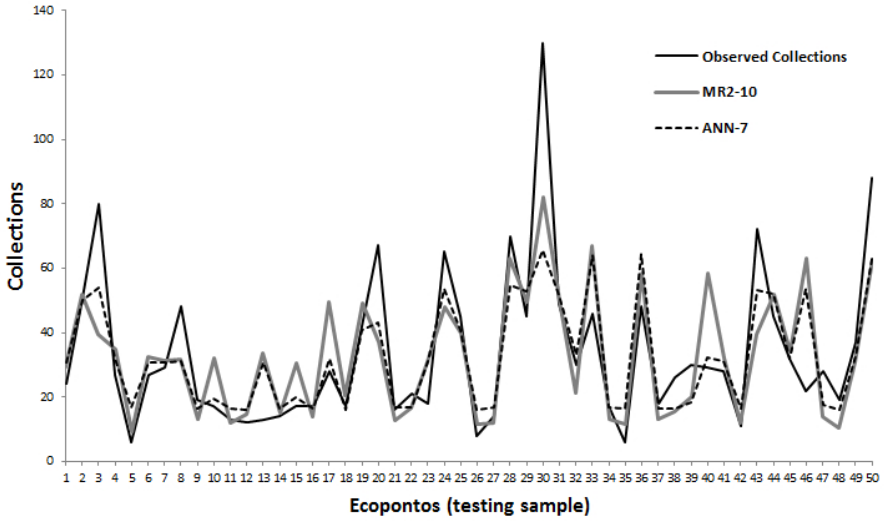


Fig. 3. Error distribution with MR2-10 model.

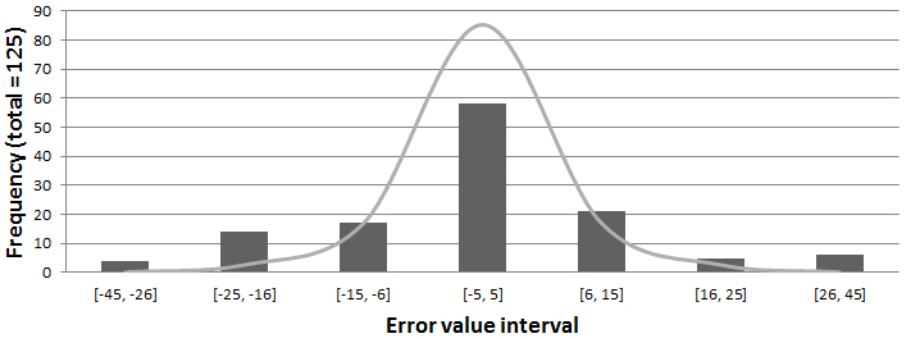
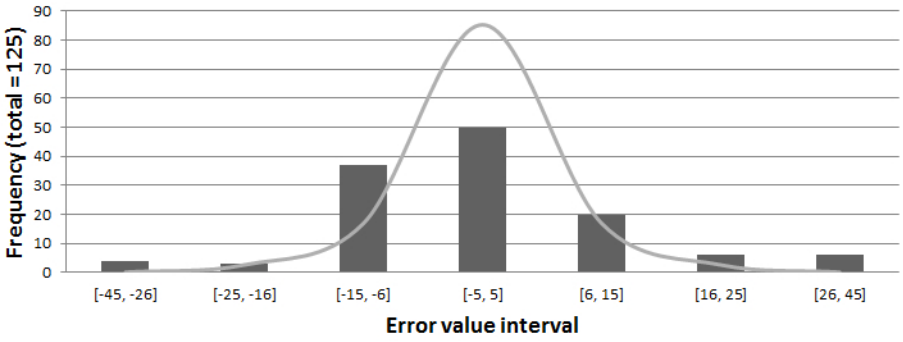


Fig. 4. Error distribution with ANN-7 model.



6 Conclusions and Future Work

In this study we presented a real-world problem faced by a company that collects House Packaging Waste (HPW) for recycling stored along a network of waste collection points (ecopontos). Our main goal was to develop a forecasting model in order to predict the number of waste collections per year for each ecoponto in the network. Other purpose of this study was to evaluate which are the most important factors for HPW generation. To accomplish that, we developed Multiple Linear Regression (MLR) and Artificial Neural Network (ANN) models. The two methodologies were tested and compared against each other. The best models developed with each methodology achieved similar performances, with a slight advantage for the ANN model, with R^2 equal to 0.672 and MAD of 9.14.

Both ANN and regression models failed to explain around 35% of the dependent variable CPY (32.8% ANN and 36.4% MLR), and that may be due to a random component that we were not able to unveil, probably because the available data did not included the quantities of waste collected at each ecoponto, and since an ecoponto can be collected when it is only half-full, the number of yearly collections we predicted may not be accurately related to waste generation rate (considering completely filled ecoponto). This missing data could help at determining a more accurate waste generation rate, based on the exact waste quantities collected instead of the number of collections made. In addition, some of the factors, such as DF and SEF, can be affected by human-error.

For future work, since MLR and ANN models have similar performances but are very different from each other in terms of methodology, we intend to combine the predictions from both models trying to improve prediction accuracy. Other hypothesis could also be tested to achieve better results for both MLR and ANN models, for example:

- Add more detail to the SEF sub-factors (check incompleteness).
- Test different ways of determining SEF.

– To correct the estimates of CPY for seasonality.

Regarding the future applications, it is our intention to label each ecoponto with a collection priority level, based on their HPW generation rate, which can be determined using the predictions of yearly waste disposals. Other relevant use for the forecasting models would be to analyse possible new locations for ecopontos along an existent network, or when designing a new collection network.

Acknowledgments This work has been supported by FCT – Fundação para a Ciência e Tecnologia within the Project Scope: PEst-OE/EEI/UI0319/2014.

References

1. Bandara N.J.G.J., Hettiaratchi J.P.A., Wirasinghe S.C., and Pilapiiya, S.: Relation of waste generation and composition to socio-economic factors: a case study. Environmental Monitoring and Assessment 135(1–3), 31–39 (2007)
2. Beigl, P., Lebersorger, S., and Salhofer, S.: Modelling municipal solid waste generation: A review. Waste Management 28(1), 200–214 (2008)
3. Bruvoll, A., and Spurkland, G.: Waste in Norway up to 2010, reports 95/8. Statistics Norway (1995)
4. Coopers and Lybrand: Cost-Benefit Analysis of the Different Municipal Solid Waste Management Systems: Objectives and Instruments for the Year 2000. Brussels, Belgium: European Commission DG XI, final report (1996)
5. Grossman, D., Hudson, J.F., and Mark, D.H.: Waste generation methods for solid waste collection. Journal of Environmental Engineering-ASCE 6, 1219–1230 (1974)
6. Marquez, M.Y., Ojeda S., and Hidalgo, H.: Identification of behavior patterns in household solid waste generation in Mexicali's city: Study case. Resources, Conservation and Recycling52(11), 1299–1306 (2008)
7. Nagelhout, D., Joosten, M., and Wierenga, K.: Future waste disposal in the Netherlands. Resources, Conservation and Recycling 4(4), 283–295 (1990)
8. Noori, R., Abdoli, M.A., Jalili Ghazi Zade, M., Samieifard, R.: Comparison of Neural Network and Principal Component Regression Analysis to Predict the Solid Waste Generation in Tehran. Iranian J Publ Health 38(1), 74–84 (2009)
9. Pham, D., and Liu, X.: Neural Networks for Identification, Prediction and Control. Berlin, Germany: Springer-Verlag (1995)
10. Shahabi, H., Khezri, S., Ahmad, B.B. and Zabihi, H.: Application of Artificial Neural Network in Prediction of Municipal Solid Waste Generation (Case Study: Saqqez City in Kurdistan Province). World Applied Sciences Journal 20 (2), 336–343 (2012)
11. Skovgaard M, Hedal N and Villanueva A.: Municipal Waste Management and Greenhouse Gases. European Topic Centre on Resource and Waste Management, working paper 2008/1. In: European Topic Centre on Sustainable Consumption and Production, Publications. (2008)
12. Wang, H.T. and Nie, Y.F.: Municipal solid waste characteristics and management in China. Journal of the Air and Waste Management Association 51(2), 250–263 (2001)
13. Zade, M. and Noori, R.: Prediction of Municipal Solid Waste Generation by Use of Artificial Neural Network: A Case Study of Mashhad. Int. J. Environ. Res. 2(1), 13–22 (2008)