Global Journal of Environmental Science and Management
(GJESM)

Homepage: https://www.gjesm.net/

ORIGINAL RESEARCH PAPER

Forecast generation model of municipal solid waste using multiple linear regression

J.A. Araiza-Aguilar[1],*, M.N. Rojas-Valencia[2], R.A. Aguilar-Vera[3]

[1]School of Environmental Engineering, University of Science and Arts of Chiapas, North beltway, Lajas Maciel, Tuxtla Gutierrez, Chiapas, Mexico

[2]Institute of Engineering, National Autonomous University of Mexico, External circuit, University City, Coyoacan delegation, Mexico City, Mexico

[3]Institute of Geography, National Autonomous University of Mexico, External circuit, University City, Coyoacan delegation, Mexico City, Mexico

ARTICLE INFO

Article History:
Received 20 April 2019
Revised 11 September 2019
Accepted 12 October 2019

Keywords:
Explanatory variables
Forecast model
Multiple linear regression
Statistical analysis
Waste generation

ABSTRACT

The objective of this study was to develop a forecast model to determine the rate of generation of municipal solid waste in the municipalities of the Cuenca del Cañón del Sumidero, Chiapas, Mexico. Multiple linear regression was used with social and demographic explanatory variables. The compiled database consisted of 9 variables with 118 specific data per variable, which were analyzed using a multicollinearity test to select the most important ones. Initially, different regression models were generated, but only 2 of them were considered useful, because they used few predictors that were statistically significant. The most important variables to predict the rate of waste generation in the study area were the population of each municipality, the migration and the population density. Although other variables, such as daily per capita income and average schooling are very important, they do not seem to have an effect on the response variable in this study. The model with the highest parsimony resulted in an adjusted coefficient of 0.975, an average absolute percentage error of 7.70, an average absolute deviation of 0.16 and an average root square error of 0.19, showing a high influence on the phenomenon studied and a good predictive capacity.

DOI: 10.22034/gjesm.2020.01.01

NUMBER OF REFERENCES	NUMBER OF FIGURES	NUMBER OF TABLES
34	5	6

*Corresponding Author:
Email: *araiza0010@hotmail.com*
Phone: +52 961176 8673
Fax: +52 96161 70440

Note: Discussion period for this manuscript open until April 1, 2020 on GJESM website at the “Show Article.

INTRODUCTION

Because of its high management cost, the amount of Municipal Solid Waste (MSW) generated in population settlements is a significant factor for the provision of public services. According to Intharathirat *et al.* (2015); Keser *et al.* (2012); Khan *et al.* (2016), the amount of MSW and its composition vary depending on social, environmental and demographic factors. Several researchers have developed models to predict the amount of MSW generated (Mahmood *et al.,* 2018; Kannangara *et al.,* 2018; Pan *et al.,* 2019; Soni *et al.,* 2019), while others analyze the variables that influence their generation and composition (Chhay *et al.,* 2018; Grazhdani, 2016; Liu and Wu, 2010; Liu *et al.,* 2019; Rybová *et al.,* 2018). Unfortunately, due to the social, economic and geographical heterogeneity of the different regions of the world, it is difficult to make inferences or projections with the proposed models, and therefore, the models and their variables have to be adapted to the conditions of other regions, sometimes with little success. Kumar and Samandder (2017) and Shan (2010) reports that some of the difficulties for the adaptation of these models are related to limited or inaccessible information in other countries (databases). In addition, some variables are theoretically valid, but difficult to measure. In other cases, the variables used do not provide information leading to the explanation of the phenomenon, but have to be used, because the model incorporates them. Mexico, this topic has also been addressed, particularly in the center and north of the country (Buenrostro *et al.,* 2001; Márquez *et al.,* 2008; Ojeda *et al.,* 2008; Rodríguez, 2004). However, it is evident that the models proposed are not applicable to the entire national context. According to the OECD (2015), there are notable differences between the central, northern and especially southern regions of Mexico; these include disparities in income, education, access to services, dispersion of localities and other factors, which cause that the consumption patterns, and therefore the amount of MSW, vary greatly. This study presents a model to forecast the generation rate of MSW in the municipalities of the Cuenca del Cañón del Sumidero (CCS), Chiapas State, Mexico. The model considers the information of the most relevant and easily accessible social and demographic variables for the study area, which correspond to statistical data for the years 2010-2015. This model will allow the decision makers of the municipalities of the CCS to determine the quantities of MSW generated, operate properly the waste management systems, and even acquire infrastructure. This study has been carried out in municipalities of the Cuenca del Cañón del Sumidero, Chiapas State, Mexico during 2010 - 2015.

MATERIALS AND METHODS

Description of the study area and context

The CCS is located in the State of Chiapas, in the southeast of Mexico, between the coordinates 15° 56’ 55” and 16° 57’ 26” North Latitude, and 92° 30’ 44” and 93° 44° 35” west longitude (Fig. 1). The CCS has 24 municipalities and 2,847 localities; 2,816 localities are rural while 31 are urban. 83% of the population of the study area lives in urban areas (INEGI, 2010). The degree of dispersion is high, especially in the rural localities farthest from the municipal seat.

Development of the model

This study uses a multiple linear regression (MLR) model to obtain the generation rates of MSW. Because of their versatility and well-founded theory, MLR models have been widely used in various scientific fields. Their main disadvantage is the preparation of the database (Pires *et al.,* 2008). The hypothesis to use the MLR in this study is based on the effect of the explanatory variables (social and demographic variables) on the response variable (generation rate of MSW). The linear function is shown in Eq. 1.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k + \varepsilon \qquad (1)$$

Where, *Y* is the response variable, X_i (1, 2, 3 ... k) are the explanatory variables, β_i (1, 2, 3 ... k) are the regression coefficients and ε is the residual error.

According to Agirre (2006), the MLR is based on two assumptions: i) the explanatory variables must be independent, i.e., free of multicollinearity and ii) the dependent variable must be normally distributed, with zero mean and constant variance. In order to determine the regression coefficients, the least squares method, which is based on minimizing the sum of squared errors (SSE), using Eq. 2.

$$SSE = \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i \right)^2 \qquad (2)$$

Where Y_i is the value of each observation and $\hat{Y}_i$ is the predicted value. Theoretically, low *SSE* values reflect a better fit of the regression model

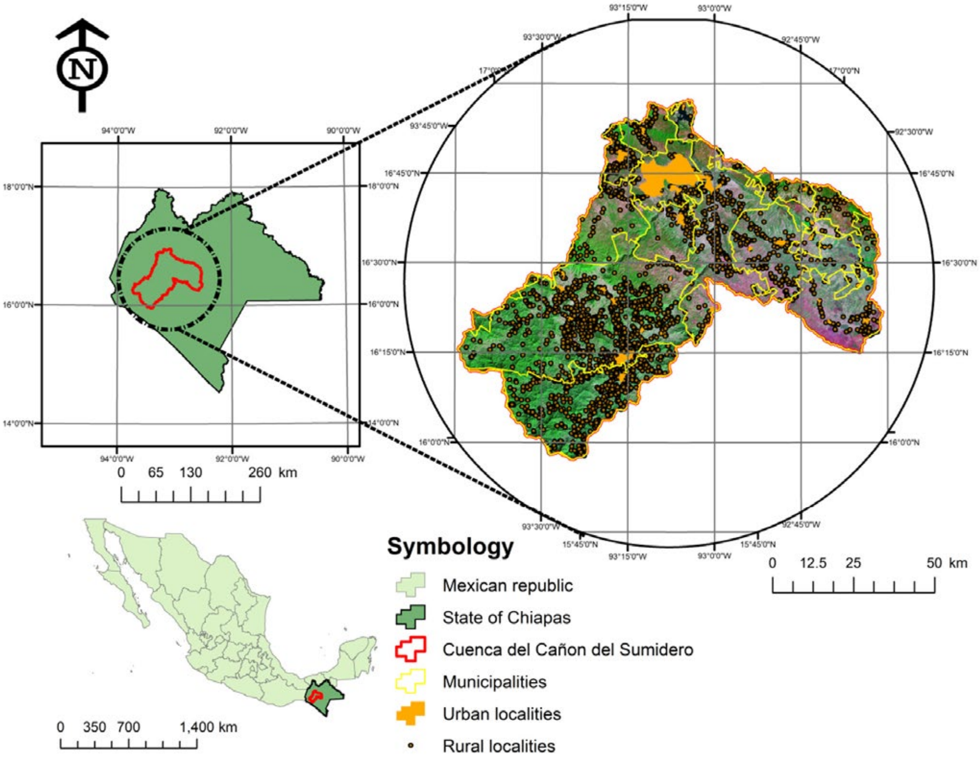

Fig. 1: Geographic location of the study area in Cuenca del Cañón de Sumidero in Mexico

(Kumar and Samandder, 2017). In order to determine the best regression model (most parsimony), the statistical significance of the explanatory variables and the general model were analyzed. The analysis of the explanatory variables was performed with the t-test, while the degree of adjustment and usefulness of the proposed model was performed by evaluating the F-test and the value of R^2_{adj} using Eqs. 3 and 4, respectively.

$$F = \frac{\left(SS_{YY} - SSE\right)/k}{SSE/\left[n-\left(k+1\right)\right]} \tag{3}$$

$$R^2_{adj} = 1 - \left[\frac{\left(n-1\right)}{n-\left(k+1\right)}\right]\left(1-R^2\right) \tag{4}$$

Where $SS_{YY} = \Sigma\left(Y_i - \bar{Y}\right)^2$ represents the sum of the squares of the difference of the observed data (Y_i) and the average of the data ($\bar{Y}$); *k* is the number of explanatory variables included in the model; *n* is the sample size; and R^2 is the coefficient of determination. The value of R^2 was not considered to measure the explanatory power of the regression model, because its value increases when adding more explanatory variables, and it can be a deceptive measure (Chang *et al.,* 2007).

Data collection

According to Beigl *et al.* (2008) and Kolekar *et al.* (2016), the methods of data collection depend on the scale of the study. In investigations carried out at household or locality levels, the acquisition of information is usually carried out through surveys or interviews; while at district or country scales, the information comes from a database registered by government agencies. This study was made at district scale and therefore the study area includes several municipalities. MSW generation was obtained from SEMANH (2013), the studies by Alvarado *et al.* (2009) and Araiza *et al.* (2015). The social and demographic information (explanatory variables) was obtained from CONAPO (2017) and INEGI (2010). The compiled information allowed the elaboration of a database of 9 variables, with 118 specific data per variable, coming from all the municipalities of the state of Chiapas (Table 1). The inferences of the proposed model were made on the municipalities of the CCS. This database was analyzed with the MINITAB software version 16.

Exploratory analysis of variables

An exploratory analysis of the 9 variables used to check the normality of the data was carried out. The test used was Kolmogorov-Smirnov, with a level of significance of $\propto = 0.05$. This test showed that the variables Y_{Gen}, X_{Pop}, X_{Pd}, X_{Pbam}, X_{Hgs}, X_{Ces}, X_{Dpi}, did not follow a normal distribution, because their p-value was smaller than the $\propto$ value considered. In order to adjust their values, the variables were transformed with natural logarithms. The variables X_{As} and X_{Mi} were not transformed because their data followed a normal distribution (Table 2).

Multicollinearity analysis and variable screening

An analysis of the explanatory variables was made prior to the selection of the best MLR model. Through a multicollinearity test, some of the variables initially considered were eliminated. Especially, the variance inflation factor (VIF) and the Pearson correlation coefficient (*r*) were used. Similar to Keser *et al.* (2012), the *r* coefficient was used to detect the bivariate association, while the VIF was used to detect the multivariate correlation. Eqs. 5 and 6 describe the tests used.

$$VIF_k = \frac{1}{\left(1 - R_k^2\right)} \qquad (5)$$

$$r = \sqrt{1 - \frac{SSE}{SS_{YY}}} \qquad (6)$$

VIF value is calculated using the R^2 of the regression equation; the explanatory variables denoted by *k* are analyzed as dependent variables, while the others are used as independent variables; thus, VIF is calculated for each explanatory variable *k*.

The cut-off value of VIF used in this study was 4. According to Ghinea *et al.* (2016), when VIF < 1, the explanatory variables are not correlated; when 1 < VIF < 5, the explanatory variables are slightly correlated; and when VIF > 5 or 10, the explanatory variables are highly correlated. The value of *r* indicates the relationship between two variables (positive or

Table 1: Description of variables

No.	Name of the variable	Symbol	Type of variable	Measure
1	MSW generation	Y_{Gen}	Dependent	Tons/day
2	Population	X_{Pop}	Independent	Inhabitants
3	Population density	X_{Pd}	Independent	Inhabitants/km^2
4	Population born in another municipality	X_{Pbam}	Independent	Inhabitants
5	Average schooling	X_{As}	Independent	Years of study
6	Household with goods and services	X_{Hgs}	Independent	Percent (%)
7	Commercial establishments and services	X_{Ces}	Independent	Number of establishments
8	Daily per capita income	X_{Dpi}	Independent	Mexican pesos/day
9	Marginalization index	X_{Mi}	Independent	Percent (%)

Table 2: Normality test and transformation of variables

Original variable			Transformed variable		
	Kolmogorov-Smirnov			Kolmogorov-Smirnov	
	Statistical	p-value		Statistical	p-value
Y_{Gen}	0.338	<0.010	$ln\text{-}Y_{Gen}$	0.050	>0.150
X_{Pop}	0.281	<0.010	$ln\text{-}X_{Pop}$	0.066	>0.150
X_{Pd}	0.271	<0.010	$ln\text{-}X_{Pd}$	0.039	>0.150
X_{Pbam}	0.387	<0.010	$ln\text{-}X_{Pbam}$	0.065	>0.150
X_{As}	0.053	>0.150	X_{As}	---	---
X_{Hgs}	0.183	<0.010	$ln\text{-}X_{Hgs}$	0.054	>0.150
X_{Ces}	0.349	<0.010	$ln\text{-}X_{Ces}$	0.056	>0.150
X_{Dpi}	0.117	<0.010	$ln\text{-}X_{Dpi}$	0.056	>0.150
X_{Mi}	0.054	>0.150	X_{Mi}	---	---

negative); its value ranges between -1 and 1. There are no clearly defined cut-off in the literature. Arriaza (2006) indicates that with values of *r* greater than 0.3, there may be signs of correlation, with values greater than 0.8, there are serious problems of multicollinearity. As in Grazhdani (2016), in this study it was considered that a value of $r \geq 0.6$ (positive or negative), indicates correlation between the explanatory variables. The elimination of explanatory variables was performed in an iterative procedure, i.e., the VIF values were initially determined for the 8 variables; subsequently, the variable with the highest VIF was eliminated and the next iteration with 7 variables was performed. This elimination procedure ended when a VIF cut-off value of 4 was found. Finally, other eliminations were made based on the values of *r*. Subsequently, 3 explanatory variables were used in the search stage for a better model (of greater parsimony). The first variable selected was X_{Pop}, i.e., the "total population" of each municipality, under the hypothesis that the larger the population, the greater the consumption and thus the greater the amount of MSW generated. The second explanatory variable used was X_{Pd} "population density", under the premise that dispersion patterns or agglomeration of inhabitants per unit area influences MSW generation. The third variable used was X_{Pbam} "population born in another municipality", which can be seen as migration, i.e., people who move to other places to seek better living conditions. The process of mobilization of people causes changes in consumption patterns of a new place of settlement. Other models that do not follow the principle of parsimony were also created (more than 3 explanatory variables), but they should not be used to forecast waste generation rates, since they have very low accuracy values and some of their explanatory variables are not significant.

Accuracy of the model and validation

In order to determine the accuracy of the best model found, 3 widely used measures were employed: the Mean Absolute Percentage Error (MAPE), the Mean Absolute Deviation (MAD) and the Root Mean Square Error (RMSE) (Eqs. 7, 8 and 9, respectively). A value of these measures close to zero indicates a high precision of the model (Azadi and karimí 2016).

$$MAPE = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{A_t - F_t}{A_t}\right| *100 \qquad (7)$$

$$MAD = \frac{1}{n}\sum_{t=1}^{n}\left|A_t - F_t\right| \qquad (8)$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{t=1}^{n}(A_t - F_t)^2} \qquad (9)$$

In these equations, A_t is the observed value, F_t is the predicted value and *n* is the sample size. MAPE is expressed in terms of percentage of error, MAD expresses the precision in the units of the data analyzed, and RMSE indicates how concentrated the data are around the line of best fit. In order to perform the external validation of the model, the technique called $R^2 jackknife$ using Eq. 10. This equation is calculated by systematically eliminating each observation from the data set, estimating the regression equation and determining to what extent the model is able to predict the observation that was removed.

$$R^2 jackknife = 1 - \frac{\Sigma\left(y_i - \hat{Y}_{(i)}\right)^2}{\Sigma\left(y_i - \bar{Y}\right)^2} \qquad (10)$$

The $R^2 jackknife$ coefficient varies between 0 and 100%, larger values suggesting models with greater predictive capacity; $\hat{Y}_{(i)}$ denotes the predicted value for i-th observation obtained when the regression model fits the data with y_i omitted (or removed) from the sample; and $\bar{Y}$ is the simple average of the observed data.

Verification of model assumptions

The validity of the MLR models is subject to the behavior of the residual errors "ε" (difference between observed and predicted values of the dependent variable), particularly their normal distribution, their independence and homoscedasticity (Kumar and Samandder, 2017). The verification of normality was carried out through the Kolmogorov-Smirnov test, with a level of significance of $\propto$ = 0.05. In order to verify the independence of residues, the Durbin-Watson test (d) was applied, looking for values close to 2, because "d" varies between 0 and 4 (Mendenhall and Sincich, 2012). The homoscedasticity assumption was evaluated with the plot of residuals vs predicted, both standardized, looking for a residue behavior that does not fit any known pattern.

RESULTS AND DISCUSSION

Statistical analysis of variables

The initial exploratory analysis was performed on the response variable Y_{Gen}, which has the behavior shown in Fig. 2. It is observed that some municipalities, which appear to be outliers, show a very high rate of MSW generation.

These atypical values were not eliminated from the analysis because they are not errors, but rather data that come from the most important municipalities of Chiapas, such as "Tuxtla Gutiérrez,

Comitán, San Cristóbal de las Casas and Tapachula". These municipalities are regional heads, therefore, the number of inhabitants, their patterns of consumption and MSW generation, differ significantly from the rest of the studied area. The normality test of the response variable and of the 8 explanatory variables is shown in Fig. 3. The non-normality of the variables Y_{Gen}, X_{Pop}, X_{Pd}, X_{Pbam}, X_{Hgs}, X_{Ces} and X_{Dpi}, can be seen. For this reason, these variables were transformed using natural logarithms (Fig. 3a, 3b, 3c, 3d, 3f, 3g and 3h).

Forecast model

The coefficients of the MLR model were determined using the Minitab software. Only the explanatory variables that fulfilled the multicollinearity criterion were used. Initially, 2 theoretically valid models were determined; the first one is shown in Eq. 11.

$$lnY_{Gen} = -8.91 + 1.10\, ln_{X\ Pop} + 0.0259\, ln_{X_{Pd}} + 0.0688\, ln_{X_{Pbam}} \tag{11}$$

This first model consists of 3 variables, X_{Pop}, X_{Pd}, X_{Pbam} (all transformed). The F-test associated with a variance analysis indicated that the model is statistically valid because p-value < 0.05. This model can thus also be used for forecast purposes. However, it is important to be careful because the explanatory variable X_{Pd} is not statistically significant since the null hypothesis that the coefficient of the variable is equal to zero (H_0: $\beta_i = 0$) is met. Therefore, the explanatory variable is not related to the dependent variable, i.e., it should not be interpreted.

The second model is presented in Eq. 12, which consists of 2 explanatory variables "X_{Pop} and X_{Pbam}". Similar to the first model, here also the p-value and the F-test indicate that it is a statistically valid model that can be used for forecasting purposes. Particularly this model is the one of greater parsimony, because it uses only 2 variables.

$$lnY_{Gen} = -8.86 + 1.11\, ln_{X\ Pop} + 0.0658\, ln_{X_{Pbam}} \tag{12}$$

All the information associated with the analysis of variance is presented in Table 3.

The verification of assumptions of the proposed models, especially model 2, is presented in Fig. 4. The probability-probability plot (p-p plot) (Fig. 4a) shows the values of the residuals with a linear pattern indicating normality; additionally, the Kolmogorov-Smirnov value and its associated p-value confirm it (p-value > 0.15). The result of the Durbin-Watson independence test gave a value of 1.979 for model 2, which indicates that the residuals are not correlated. The homoscedasticity test presented in

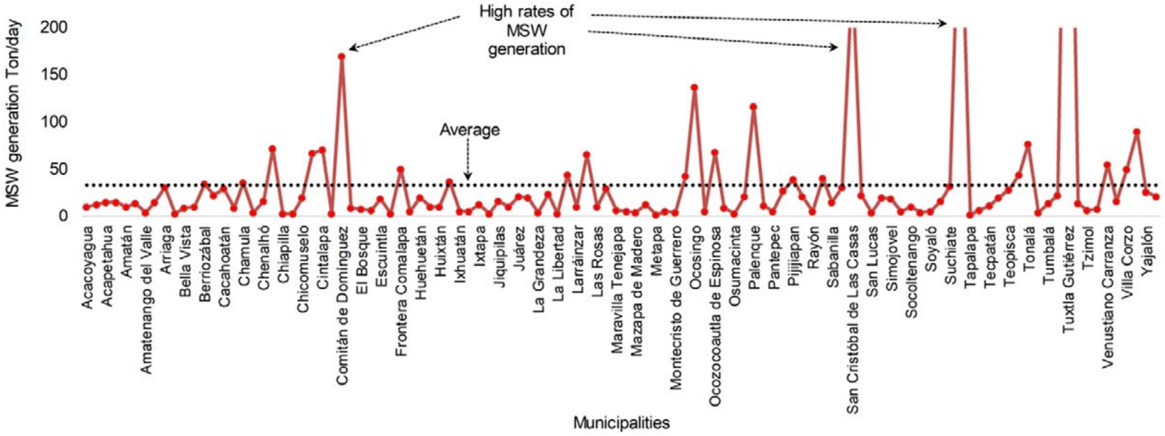

Fig. 2: MSW generation rates in the state of Chiapas

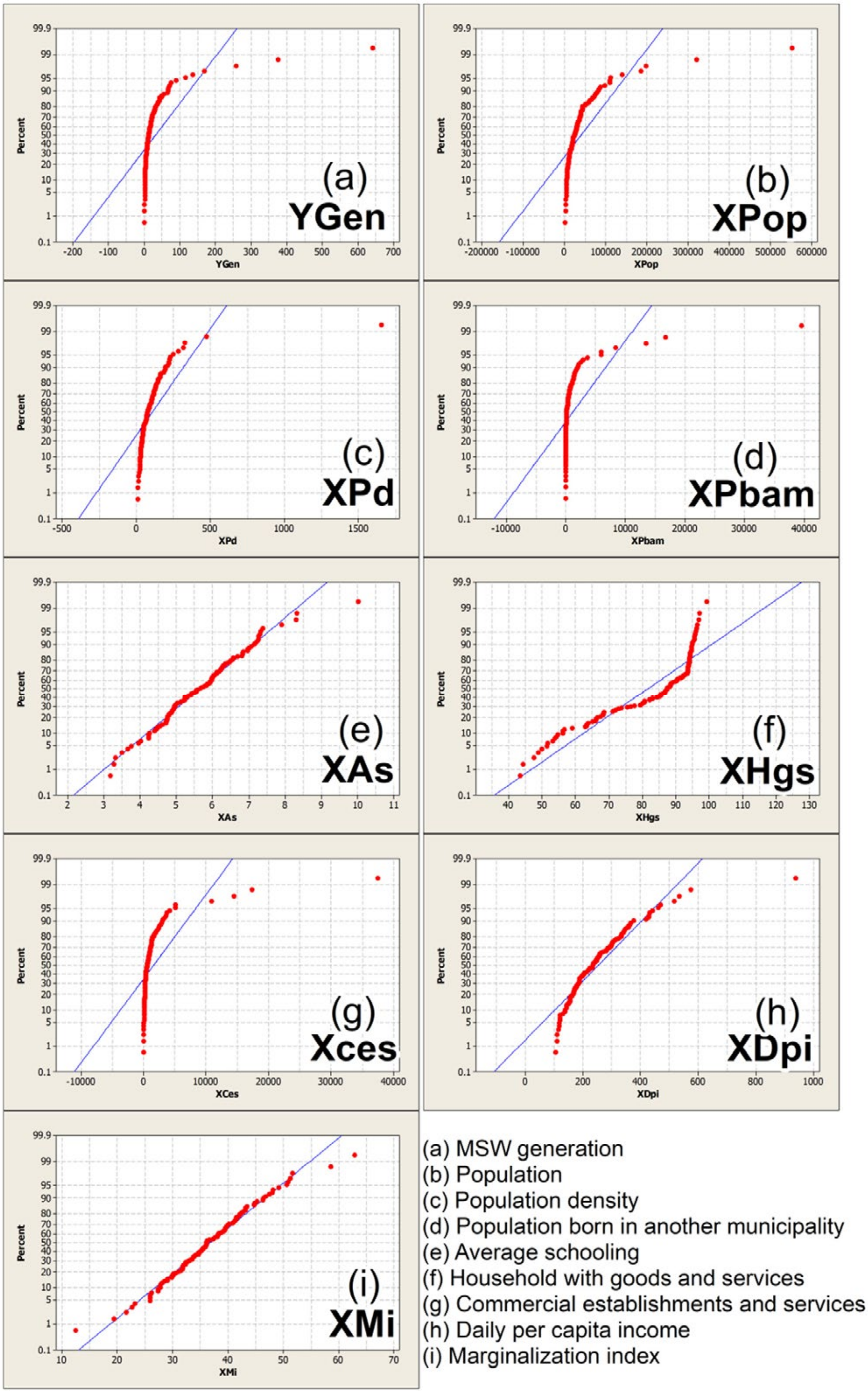

Fig. 3: Behavior of the variables analyzed with respect to normality

Table 3: Analysis of variance of the proposed models

Model	Source	Degree of freedom (df)	Sum of Squares	Mean square	F	Sig.
	Regression	3	150.591	50.197	1,449.33	0.000
1	Residual	107	3.706	0.035		
	Total	110	154.297			
	Regression	2	150.549	75.274	2,169.16	0.000
2	Residual	108	3.748	0.035		
	Total	110	154.297			

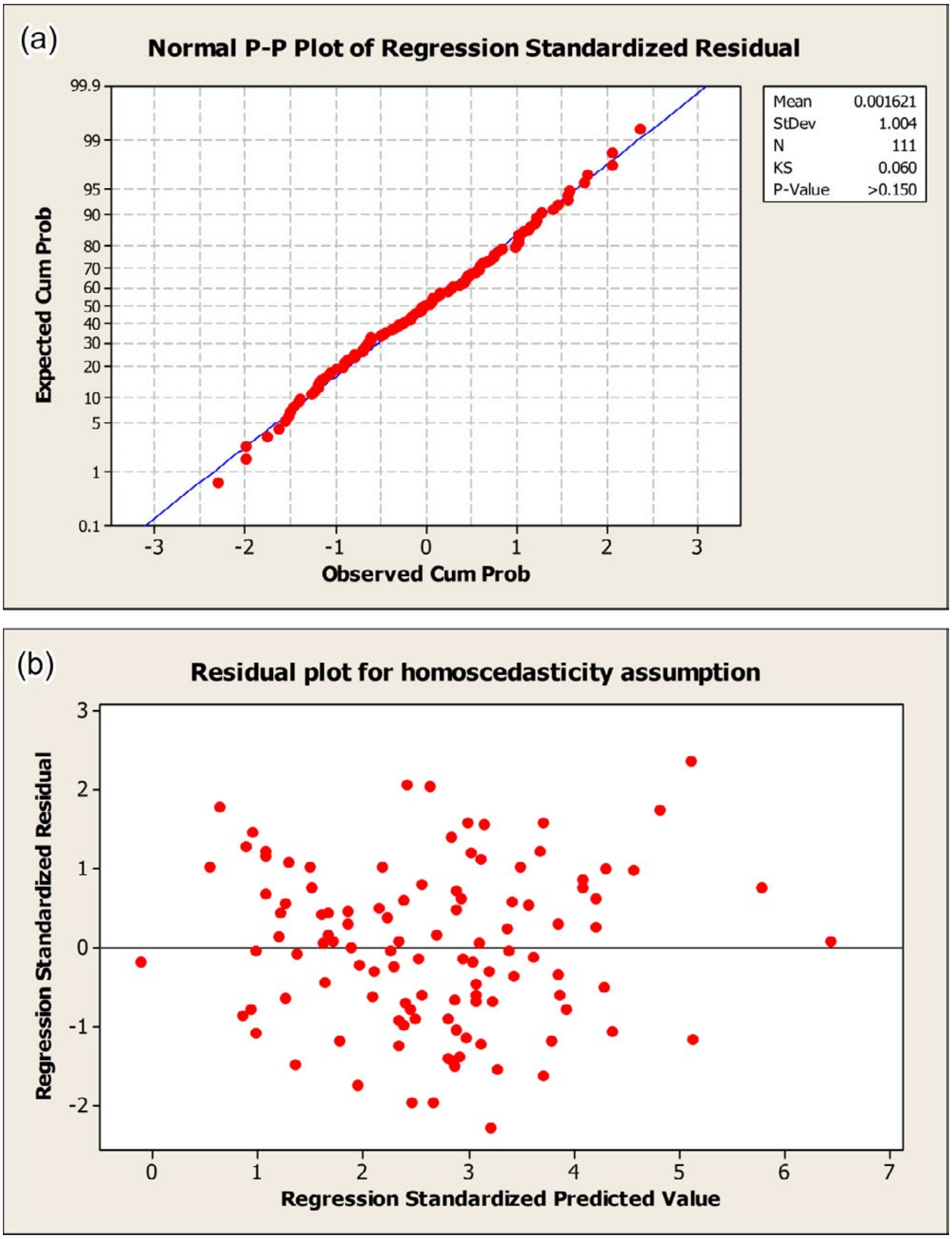

Fig. 4: Verification of model assumptions: (a) Normality of residuals, (b) Independence of residuals

Fig. 4b shows a behavior of the residuals that does not fit any known pattern; therefore, this situation is adequate.

On the other hand, the R^2 value of the equations in both models was 0.976, which indicates that 97.6% of the generation rate of MSW Y_{Gen} (transformed) can be explained by the explanatory variables used. It is important to note the gradual decrease of R^2_{adj} (0.975) with respect to R^2, which is due to the adjustment by the introduction of 2 and 3 variables in models 1 and 2, respectively. The high value of R^2 and R^2_{adj} in these models is due to the initial transformation of the explanatory variables, as well as the response variable. Additionally, the data collection carried out in this study influenced these values because they come from a census database, and not from an information survey through interviews. The internal validation of model 2 through MAPE, MAD and RMSE, showed the values of 7.70, 0.16 and 0.19, respectively, which indicates a high precision since the values of these tests are close to 0 (zero). The external validation by $R^2 jackknife$ presented a value of 97.44%. Therefore, model 2 also has a high forecasting capacity.

Non-significant variables

The analysis of the 8 explanatory variables using the VIF test produced the initial elimination of the variables X_{As}, *ln*-X_{Hgs}, *ln*-X_{Ces} and *ln*-X_{Dpi}, since their value was higher than the cut-off of 4. The variables X_{As} and *ln*-X_{Dpi} have been used mainly in studies at household or locality levels (Khan *et al*, 2016; Ojeda *et al.,* 2008), but in this paper they were used at district level, and the effect of these variables seems not to be important (low correlation with the response variable *ln*-Y_{Gen}). The variables *ln*-X_{Hgs} and *ln*-X_{Ces} were eliminated because they are highly correlated with X_{Mi}, since the latter is a multidimensional indicator that measures deprivation in a population, through variables similar to those eliminated. Finally, through the *r* test, only X_{Mi} was eliminated, since it was highly correlated with *ln*-X_{Pbam}, with a coefficient of -0.695, i.e., much higher than the cut-off value of 0.6 (positive or negative); additionally, this variable was less correlated with the *ln*-Y_{Gen} response variable (Table 4).

Significant variables

The transformed variables X_{Pop}, X_{Pd} and X_{Pbam} were used in the search for the best model, since their VIF and r values were below the cut-off values. X_{Pop} has been used in the studies of Azadi and karimí (2016) and Abdoli *et al.* (2011), as the most important explanatory variable. In this study, Pearson's correlation *r*-value was 0.985, which indicates that it is also the variable most related to the generation of waste, particularly in a positive way, i.e., to a larger population corresponds a greater quantity of MSW. The variable X_{Pd} has been used in few publications. Bel and Mur (2009) use this variable also to obtain the costs associated with waste management. In this study, *r*-value of 0.161 was obtained, which indicates a poor correlation with the response variable. The analysis of the forecast model 1 indicated that this

Table 4: Correlation matrix of variables

Pearson's correlation	*ln*-Y_{Gen}	*ln*-X_{Pop}	*ln*- X_{Pd}	*ln*- X_{Pbam}
ln-X_{Pop}	0.985	---	---	---
ln- X_{Pd}	0.161	0.169	---	---
ln- X_{Pbam}	0.638	0.573	-0.059	---
X_{Mi}	-0.355	-0.271	-0.097	-0.695

Table 5: Analysis of variance of other generated models

Model	Source	Degree of freedom (df)	Sum of Squares	Mean square	F	Sig.	R^2_{adj}
3	Regression	4	150.900	37.725	1,177.44	0.000	0.977
	Residual	106	3.396	0.032			
	Total	110	154.297				
4	Regression	5	150.902	30.180	933.42	0.000	0.977
	Residual	105	3.395	0.032			
	Total	110	154.297				
5	Regression	6	151.397	25.233	905.01	0.000	0.980
	Residual	104	2.900	0.028			
	Total	110	154.297				

variable is not statistically significant, and its use must be taken with caution. The X_{Pbam} variable is positively related to the response variable. Its Pearson's correlation coefficient was 0.638. This variable is important in the study area, since it can be concluded that people who move from one municipality to another have different consumption patterns that modify the amounts of MSW. Other explanatory variables mentioned in Kolekar *et al.* (2016), for instance age, employment status, level of urbanization and environmental variables such as precipitation or temperature, were not used in this study since it is difficult to find a database with information on these variables.

Other generated models

Eqs. 13, 14 and 15 show other models generated with the variables initially raised (models 3, 4 and 5 respectively). All these models are statistically significant and are also useful for forecasting purposes, but incorporate explanatory variables that are not significant; therefore, their results are not accurate (Table 5). Additionally, they have low parsimony because they incorporate more than 2 or 3 explanatory variables. Table 6 shows the statistical behavior of the predictors. The p-value and the VIF must be analyzed because they indicate multicollinearity between the variables and also their possible interpretation within the generated model.

$$lnY_{Gen} = -8.48 + 1.13\, ln_{X\,Pop} - 0.0019\, ln_{X_{Pd}} + 0.0315\; ln_{X_{Pbam}} - 0.0111 X_{Mi} \tag{13}$$

$$lnY_{Gen} = -8.42 + 1.13\, ln_{X\,Pop} - 0.0008\, ln_{X_{Pd}} + 0.0324\; ln_{X_{Pbam}} - 0.0070 X_{As} - 0.0119 X_{Mi} \tag{14}$$

$$lnY_{Gen} = -8.22 + 0.995\, ln_{X\,Pop} + 0.0039\, ln_{X_{Pd}} + 0.0261\; ln_{X_{Pbam}} - 0.0006 X_{As} + 0.133\, ln_{X_{Hgs}} - 0.00525 X_{Mi} \tag{15}$$

Inferences about the municipalities of the study area

Based on model 2 and its statistical analysis, inferences were made to forecast the generation rate of MSW in the municipalities of the CCS. The forecast was made with the most current data of the variables X_{Pop} and X_{Pbam}, corresponding to the year 2015. Fig. 5 shows MSW generation forecast and its comparison with the original database. In most of the municipalities of the CCS, the generation rate of MSW presented a gradual increase with respect to population growth (variable X_{Pop}), except in Arriaga, Chiapilla, Osumacinta, Suchiapa, Teopisca, Tonalá, Venustiano Carranza and Villaflores, due to the fact that the population of these municipalities did not increase in the 2010-2015 period.

Currently, the study area generates 1,600 tons of MSW/day, of which 74% comes from the regional heads such as Berriozábal, Ocozocoautla de Espinosa, San Cristóbal de las Casas, Tuxtla Gutiérrez and Villaflores.

Table 6: Statistical behavior of predictors in models 3, 4 and 5

Model	Predictor	Coefficient	p-value	VIF	Comments
3	Constant	-8.48	0.000	-------	There is no multicollinearity among predictors, but some of them are not statistically significant (p-value> 0.05).
	$ln_{X_{Pop}}$	1.13	0.000	1.947	
	$ln_{X_{Pd}}$	-0.0019	0.937	1.300	
	$ln_{X_{Pbam}}$	0.0315	0.053	3.556	
	X_{Mi}	-0.0111	0.002	2.405	
4	Constant	-8.42	0.000	-------	There is multicollinearity between predictors (VIF> 5) and some of them are not statistically significant (p-value> 0.05).
	$ln_{X_{Pop}}$	1.13	0.000	1.948	
	$ln_{X_{Pd}}$	-0.0008	0.975	1.374	
	$ln_{X_{Pbam}}$	0.0324	0.055	3.789	
	X_{As}	-0.0070	0.844	5.373	
	X_{Mi}	-0.0119	0.026	5.177	
5	Constant	-8.22	0.000	-------	
	$ln_{X_{Pop}}$	0.995	0.000	5.634	
	$ln_{X_{Pd}}$	0.0039	0.869	1.377	
	$ln_{X_{Pbam}}$	0.0261	0.096	3.824	
	X_{As}	-0.0006	0.986	5.384	
	$ln_{X_{Hgs}}$	0.133	0.000	6.180	
	X_{Mi}	-0.0525	0.309	5.710	

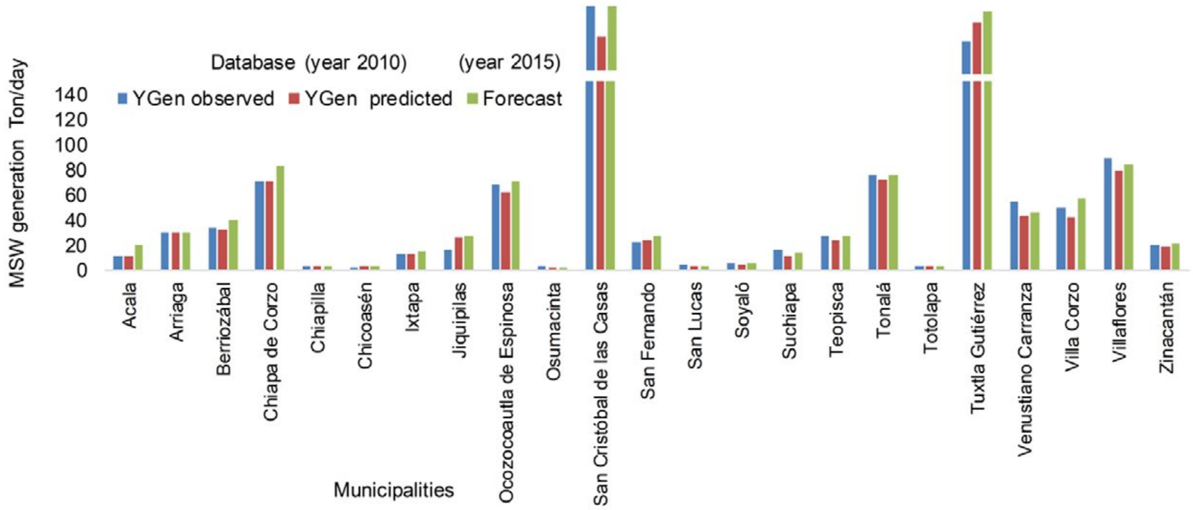

Fig. 5: Forecast of MSW generation rates in the municipalities of the study area

CONCLUSION

In this study, a forecast model was developed to determine the generation of MSW in the municipalities of the CCS, Chiapas State, Mexico. A MLR was used to obtain the forecast model with social and demographic explanatory variables. Two forecast models were presented and analyzed, with variables that met the multicollinearity test. The most important variables to predict the rate of MSW generation in the study area were the population of each municipality (X_{Pop}), the population born in another municipality (X_{Pbam}) and the population density (X_{Pd}). X_{Pop} is the most influential explanatory variable of waste generation, particularly it is related in a positive way. X_{Pbam} is less related to waste generation. X_{Pd} is the variable that least influences waste generation prediction; in addition, it can present problems of correlation with other explanatory variables. Although other variables, such as daily per capita income (X_{Dpi}) and average schooling (X_{As}), are very important, they do not seem to have an effect on the response variable in this study. The user of this forecast model should use model 2, since it is the one with the highest parsimony (it uses fewer variables); R^2_{adj}, MAPE, MAD and RMSE values indicated high influence on the explained phenomenon and high forecasting capacity. Additionally, it is important to mention that when using the models proposed for forecasting purposes, it is necessary to make a transformation in the explanatory and response variables (use inverse of natural logarithm). The inferences made on the municipalities of the study area showed that, except in some municipalities, the MSW generation rate usually presented a gradual increase with respect to population growth and with respect to the number of inhabitants that were born in another entity (migration). Finally, this study can be a solid basis for comparison for future research in the area of study. It is possible to use different mathematical models such as artificial neural network, principal component analysis, time-series analysis, etc., and compare the response variable or the predictors.

ACKNOWLEDGEMENTS

This study was supported by the Project Support Programme for Research and Technological Innovation [Project UNAM DGAPA-PAPIIT IN105516]. We also appreciate the support of Mexican National Council for Science and Technology (CONACYT) to carry out this work.

CONFLICT OF INTEREST

The author declares that there is no conflict of interests regarding the publication of this manuscript. In addition, the ethical issues, including plagiarism, informed consent, misconduct, data fabrication and/or falsification, double publication and/or submission, and redundancy have been completely observed by the authors.

ABBREVIATIONS

$\propto$	Level of significance
A_t	Observed value
β_i (1,2,3 ... k)	Regression coefficients
CCS	Cuenca del Cañón del Sumidero
d	Durbin-Watson test
Eq.	Equation
F	Fisher test
F_t	Predicted value
H_0	Null hypothesis
k	Number of explanatory variables included in the model
$ln\text{-}Y_{Gen}$	Natural logarithm of MSW generation
$ln\text{-}X_{Pop}$	Natural logarithm of population
$ln\text{-}X_{Pd}$	Natural logarithm of population density
$ln\text{-}X_{Pbam}$	Natural logarithm of population born in another municipality
$ln\text{-}X_{Hgs}$	Natural logarithm of household with goods and services
$ln\text{-}X_{Ces}$	Natural logarithm of commercial establishments and services
$ln\text{-}X_{Dpi}$	Natural logarithm of daily per capita income
MAD	Mean Absolute Deviation
MAPE	Mean absolute percentage error
MLR	Multiple Linear Regression
MSW	Municipal Solid Waste
n	Sample size
p-p plot	Probability-probability plot
p-value	Probability value
r	Pearson correlation coefficient
r-value	Pearson correlation coefficient
R^2	Coefficient of determination
R^2_{adj}	Adjusted coefficient of determination
$R^2 jackknife$	Jackknife coefficient of determination
RMSE	Root Mean Square Error
SSE	Sum of Squared Errors
SS_{YY}	Sum of the squares of the difference of (Y_i) and the $(\bar{Y})$
VIF	Variance Inflation Factor
X_i (1,2,3 ... k)	Explanatory variables
X_{As}	Average schooling
X_{Ces}	Commercial establishments and services
X_{Dpi}	Daily per capita income
X_{Hgs}	Household with goods and services
X_{Mi}	Marginalization index
X_{Pbam}	Population born in another municipality
X_{Pd}	Population density
X_{Pop}	Population
$\bar{Y}$	Average of observed data
Y_{Gen}	MSW generation
Y_i	Value of each individual observation
$\hat{Y}_i$	Predicted value

REFERENCES

Abdoli, M.; Falahnezhad, M.; Behboudian, S., (2011). Multivariate econometric approach for solid waste generation modeling: a case study of Mashhad, Iran. Environ. Eng. Sci., 28(9): 627-633 **(7 pages).**

Agirre, E.; Ibarra, G.; Madariaga, I., (2006). Regression and multilayer perceptron-based models to forecast hourly O_3 and NO_2 levels in the Bilbao area. Environ. Modell. Software. 21(4): 430–446 **(17 pages).**

Alvarado, H.; Nájera, H.; González, F.; Palacios, R., (2009). Study of generation and characterization of household solid waste in the municipal seat of Chiapa de Corzo, Chiapas, Mexico. Lacandonia J., 3: 85-92 **(8 pages).**

Araiza, J.; López, C.; Ramírez, N., (2015). Municipal solid waste management: case study in Las Margaritas, Chiapas. AIDIS J. Eng. Environ. Sci.: Res. Develop. Pract., 8(3): 299-311 **(13 pages).**

Arriaza, M., (2006). Practical guide to data analysis, junta de Andalucía, ministry of innovation, science and business, institute of agricultural research and training and fishing, Spain.

Azadi, S.; Karimí, 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 Manage., 48: 14-23 **(10 pages).**

Beigl, P.; Lebersorger, S.; Salhofer, S., (2008). Modelling municipal solid waste generation: a review. Waste Manage., 28(1): 200-214 (**15 pages**).

Bel, G.; Mur, M., (2009). Intermunicipal cooperation, privatization and waste management costs: evidence from rural municipalities. Waste Manage., 29(10): 2772–2778 (**7 pages**).

Buenrostro, O.; Bocco, G.; Vence, J., (2001). Forecasting generation of urban solid waste in developing countries - a case study in Mexico. J. Air Waste Manage., 51(1): 86-93 (**8 pages**).

Chang, Y.; Lin, C.; Chyan, J.; Chen, I.; Chang, J. (2007). Multiple regression models for the lower heating value of municipal solid waste in Taiwan. J. Environ. Manage., 85(4): 891–899 (**9 pages**).

Chhay, L.; Reyad, M.; Suy, R.; Islam, M.; Mian M., (2018). Municipal solid waste generation in China: influencing factor analysis and multi-model forecasting. J. Mater. Cycles Waste Manage., 20(3): 1761–1770 (**10 pages**).

CONAPO, (2017). National Population Council. Municipal Marginalization Index.

Ghinea, C.; Niculina, E.; Comanita, E.; Gavrilescu, M.; Campean, T.; Curteanu, S.; Gavrilescu, M. (2016). Forecasting municipal solid waste generation using prognostic tools and regression analysis. J. Environ. Manage., 182: 80-93 (**14 pages**).

Grazhdani, D., (2016). Assessing the variables affecting on the rate of solid waste generation and recycling: an empirical analysis in Prespa Park. Waste Manage., 48: 3-13 (**11 pages**).

INEGI, (2010). Population and housing census 2010: interactive data query. National Institute of Statistic and Geography.

Intharathirat, R.; Salam, P.; Kumar, S.; Untong, A., (2015). Forecasting of municipal solid waste quantity in a developing country using multivariate grey model. Waste Manage., 39: 3–14 (**12 pages**).

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 Manage., 74: 3–15 (**13 pages**).

Khan, D.; Kumar, A.; Samadder, S., (2016). Impact of socioeconomic status on municipal solid waste generation rate. Waste Manage., 49: 15-25 (**11 pages**).

Keser, S.; Duzgun, S.; Aksoy, A., (2012). Application of spatial and non-spatial data analysis in determination of the factors that impact municipal solid waste generation rates in Turkey. Waste Manage., 32(3): 359–371 (**13 pages**).

Kolekar, K.; Hazra, T.; Chakrabarty, S., (2016). A review on prediction of municipal solid waste generation models. Procedia Environ Sci., 35: 238 – 244 (**7 pages**).

Kumar, A.; Samandder, S., (2017). An empirical model for prediction of household solid waste generation rate – a case study of Dhanbad, India. Waste Manage., 68: 3-15 (**13 pages**).

Liu, C.; Wu, X., (2010). Factors influencing municipal solid waste generation in China: a multiple statistical analysis study. Waste Manage. Res., 29(4): 371-378 (**8 pages**).

Liu, J.; Li, Q.; Gu, W.; Wang, C., *(*2019). The Impact of consumption patterns on the generation of municipal solid waste in China: evidences from provincial data. Int. J. Environ. Res. Public Health, 16(10): 1-19 (**19 pages**).

Mahmood, S.; Sharif, F.; Rahman, A.U.; Khan, A.U., *(*2018). Analysis and forecasting of municipal solid waste in Nankana City using geo-spatial techniques. Environ. Monit. Assess., 190(5): 1-14 (**14 pages**).

Márquez, M.; Ojeda, S.; Hidalgo, H., (2008). Identification of behavior patterns in household solid waste generation in Mexicali's City: study case. Resour. Conserv. Recycl., 52(11): 1299–1306 (**8 pages**).

Mendenhall, W.; Sincich, T., (2012). A second course in statistics regression analysis, Prentice Hall, United States of America.

OECD, (2015). Measuring well-being in Mexican States. Organization for Economic Cooperation and Development. OECD Publishing, Paris, France.

Ojeda, S.; Lozano, G.; Morelos, R.; Armijo, C., (2008). Mathematical modeling to predict residential solid waste generation. Waste Manage., 28(1): S7–S13 (**7 pages**).

Pan, A.; Yu, L.; Yang, Q, *(*2019). Characteristics and forecasting of municipal solid waste generation in China. Sustainability, 11(5): 1-11 (**11 pages**).

Pires, J.; Martins, F.; Sousa, S.; Alvim, M.; Pereira, M. (2008). Selection and validation of parameters in multiple linear and principal component regressions. Environ. Modell. Softw., 23(1): 50-55 (**6 pages**).

Rodríguez, M. (2004). Design of a mathematical model of municipal solid waste generation in Nicolás Romero, Mexico. Master's Thesis, National Polytechnic Institute, Mexico.

Rybová, K.; Slavik, J.; Burcin, B.; Soukopová, J.; Kučera, T.; Černíková, A., (2018). Socio-demographic determinants of municipal waste generation: case study of the Czech Republic. J. Mater. Cycles Waste Manage., 20(3): 1884–1891 (**8 pages**).

Shan, C., (2010). Projecting municipal solid waste: the case of Hong Kong SAR. Resour. Conserv. Recycl., 54(11): 759–768 (**10 pages**).

Soni, U.; Roy, A.; Verma, A.; Jain, V., *(*2019). Forecasting municipal solid waste generation using artificial intelligence models—a case study in India. SN Appl. Sci., 1(2): 1-11 (**11 pages**).

AUTHOR (S) BIOSKETCHES

Araiza-Aguilar, J.A., Ph.D., Professor, School of Environmental Engineering, University of Science and Arts of Chiapas, North beltway, Lajas Maciel, Tuxtla Gutierrez, Chiapas, Mexico. Email: *araiza0010@hotmail.com*

Rojas-Valencia, M.N., Ph.D., Professor, Institute of Engineering, National Autonomous University of Mexico, External circuit, University City, Coyoacan delegation, Mexico City, Mexico. Email: *nrov@pumas.iingen.unam.mx*

Aguilar-Vera, R.A., Ph.D. Candidate, Institute of Geography, National Autonomous University of Mexico, External circuit, University City, Coyoacan delegation, Mexico City, Mexico. Email: *rodrigoantonioaguilarvera@gmail.com*

HOW TO CITE THIS ARTICLE

Araiza-Aguilar, J.A.; Rojas-Valencia, M.N.; Aguilar-Vera, R.A., (2020). Forecast generation model of municipal solid waste using multiple linear regression. Global J. Environ. Sci. Manage., 6(1): 1-14.

DOI: 10.22034/gjesm.2020.01.01

url: https://www.gjesm.net/article_36862.html