International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 09 Issue: 12 | Dec 2022 www.irjet.net p-ISSN: 2395-0072

Forecasting Municipal Solid Waste Generation Using a Multiple Linear Regression Model

Rayaan Kakad[1]

[1] *Jamnabai Narsee International School*

Abstract - *The prime objective of this research is to develop an accurate and simplified model to forecast municipal solid waste generation. Because if the municipality can predict Municipal Solid Waste (MSW), then necessary steps can be taken to process it, and better waste management can be done. Further, a drastic increase in the waste can cause mismanagement and severe health complications. Various machine learning methods are available for predicting dependent variables from a single or bunch of independent variables. In many cases (like non-linear dependence), pre-processing of the available data has to be performed to increase the accuracy and reliability of the linear model. This is also the case with the prediction of solid waste generation.*

In this work, the multiple linear regression method is used as it tends to lower overfitting and gives a good level of generalization. Pandas data frame is used to store and perform different operations on data. Also, different linearity checks like multi-collinearity, homoscedasticity, and homogeneity are performed for the suitability of the linear model. The present work considers solid waste generation in different municipalities of Italy. Different errors like MAE, MSA & RMSE are calculated for predicting MSW and the necessary steps that are required. Further accuracy of the model is calculated for the reliability of the results. Such work can be helpful for any municipal body around the world for better waste management through the prediction of solid waste generation.

Key Words: Linear Regression, Solid waste generation, Population

1. INTRODUCTION

Managing and disposing of waste is one of the most significant environmental challenges of the 21st century for any municipal body. To create a waste disposal plan, the primary input that one must have beforehand is the amount of waste to be disposed of. Waste can be broadly classified as solid, liquid, organic waste, and recyclable rubbish like food and beverage containers, dry newspapers, etc. [1]. Out of these four major types of waste, this study is focused on the amount of solid waste generated.

There can be multiple sources of solid waste in a municipal region, and hence different factors must be considered while predicting the overall amount produced. Solid waste management in a municipal region involves high costs and efforts from the government. Furthermore, proper waste disposal is the prime requirement to keep the environment clean and people healthy. Hence, this research aims to predict municipal solid waste (MSW) generation. Various studies in literature, like Navarro-Esbrı' et al. [2], suggest that MSW generation is a dynamic process and depends on multiple factors. So, to predict the same, a time series model is used for prediction.

Further time series modeling is used along with fuzzy logic, which is an application of an adaptive neuro-fuzzy interface system (ANFIS). ANFIS model is an improvement over the traditional autoregressive moving average (ARMA) model. Beigl et al. [3] developed a procedural guideline for crucial design options with impact significance and cost efficiency of the waste generation model.

An algorithm like the tree was developed for methodology selection where inputs like total MSW, collection stream, and material stream give different method output as regression method or time series analysis, or group comparison and regression. Their work aims at providing only one methodology for selecting the best-suited method for solid waste prediction. Pires J. [4] aimed at selecting statistically valid parameters using multiple linear regression and principal component analysis. From independent variables like temperature, relative humidity, wind speed, wind direction, and solar radiation, the dependent variable (concentration of tropospheric ozone) was derived. Estay-Ossandon, C [5] used the fuzzy Technique for Order Preference by Similarity to Ideal Solution (TOPSIS) for solid waste forecasting of the Canary archipelago. The authors used Delphi technique to capture experts' knowledge, and subsequently, TOPSIS was applied. The model developed from this method was used to forecast MSW generation for 2015-2030.

From the literature, it is deduced that using a non-linear forecasting model (like the one developed by Estay-Ossandon) not only increases the complexity of the model but also makes it challenging to gather raw data according to the requirement of different models. Also, a slight

change in the underlying data structure makes it difficult to modify the existing model. The model developed by Esbrı' et al. [2] explains the dependent variable more practically, but the research was explicitly aimed at forecasting river flow.

Hence, this work focuses on developing a simple yet powerful model that is more generalized and can be used for any municipal body with minimal modifications. Data used for the present research was collected by "Centro Studi Dipartimentale in Economia e Regolazione di Servizi, dell'Industria e del Settore Pubblico (CESISP)" [6] research center of the Department of Business and Law of the University of Milano-Bicocca, Italy. The popular research in this domain utilizes highly complex models like TOPSIS, ANN (Artificial Neural Network), LSTM (Long Short-term Memory), and fuzzy logic. These models are found to be overfitted and need more generalization. So, the present study aims at a multiple linear regression model, which primarily avoids overfitting, and therefore, modeling of the same is quick and reliable. Different linearity checks are explained and performed along with outlier detection using a boxplot. Non-linear dependencies are converted to linear models by using a logarithm. Multi-variable linear regression modeling is done using sklearn module of python, and finally, different errors are calculated for comparison with other methods.

2. TOOLS & METHODS

2.1 Model Development and Performance Evaluation

In order to predict dependent variables from independent variables, there are multiple machine learning methods available, like linear regression, decision trees, and random forest. Amongst these methods, linear regression is chosen in this work to give better generalization and simplification of the model. Multiple linear regression (MLR) can be the best model for this domain because multiple factors affect the quantity of final waste [7]. It can be modeled as below,

$$y = a_1x_1 + a_2x_2 + \cdots + a_nx_n + E$$

Here, y is the dependent variable, which depends on the independent variables $x_1, x_2, \dots, x_n$. While $a_1, a_2, \dots, a_n$ are regression co-efficient, which are to be calculated. E is an error or noise which cannot be derived or considered in the calculation as it is purely on a chance basis.

While selecting a linear model, the following assumptions need to be made.

- **Linearity:** It states that there exists a linear relationship between dependent and independent variables. In this research, exploratory data analysis (EDA) is performed and it is found that municipal waste is linearly dependent or the logarithm of the dependent variable is linearly dependent on independent variables, hence this assumption holds true.
- **Multi-collinearity:** there should not be any relationship between independent variables itself. It is checked by plotting pair plots between independent variables.
- **Homoscedasticity:** the error term must showcase constant variance. During the EDA process, a residual plot can be plotted to check this assumption.
- **Exogeneity:** It states that the error term must not be a function of independent variables ($x_1, x_2, \dots, x_n$), this also can be confirmed by the residual plot, which shows no relation between the error term and different features.

Process followed for regression model from data collection to model evaluation, is as below,

2.2 Performance Evaluation Indicators

- **Mean Absolute Error (MAE):** It is the absolute submission of the difference between the predicted value and actual value divided by the total number of observations. It y_p is the predicted value, and n is the total number of observations then formula for the same is as below. It uses same weight for both small and significant errors.

$$MAE = \frac{\sum_{i=0}^{n} | y_p - y |}{n}$$

- **Mean Squared Error (MSE):** It is the submission of the squared difference between the predicted value and actual value divided by the total number of observations. It gives high importance to larger values due to the square term. If y_p is the predicted value, and n is the total number of observations; the formula for the same is below. It puts a high weight on higher errors.

$$MSE = \frac{\sum_{i=0}^{n} (y_p - y)^2}{n}$$

- **Root Mean Squared Error (RMSE):** It gives high importance to larger value due to the square term.

If y_p is the predicted value, and n is the total number of observations, then, a formula for the same is as below. It puts higher weight on more significant errors and gives errors in the same unit as the original. Its mathematical expression is given below:

$$RMSE = \sqrt{\frac{\sum_{i=0}^{n} (y_p - y)^2}{n}}$$

All three performance matrices will be used to evaluate the model's performance [9]. The absolute accuracy of the model can be derived by the R squared value, which gives accuracy in percentage terms. The formula for the same is as below,

- **R-Squared (R^2):** It is the proportion of the variation in the dependent variable that is predictable from the independent variable. It is also called the accuracy of the model. If RSS is the sum of squares of residuals and TSS is the total sum of squares, then R2 can be written as below,

$$R^2 = 1 - \frac{RSS}{TSS}$$

2.3 Data Collection and EDA

The data collection method depends on factors like the type, scale, and the area of study. It is generally obtained from information registered by different government agencies or municipalities.

The data used here is based on a study conducted by the Department of Business and Law of the University of Milano-Bicocca [6] for different cities of Italy. Out of the different columns available in the dataset, the variance inflation factor (VIF) is calculated for all columns, and the chosen columns with a high VIF (higher than 10) are mentioned below.

Data considered for the present research has the following features (Independent variables):

1. Population
2. Population density
3. Average education
4. Houses with in-house services
5. Income per capita

From these dependent variables, the independent variable of municipal solid waste has to be predicted as the dependent variable might be linearly or non-linearly dependent on independent variables [8]. The explanation and units for each variable are mentioned in table 1.

Table 1: Data set columns explanation

Variable	Unit	Type	Explanation
Population	Inhabitants	Independent	Total number of populations in considered city or village
Population Density	Inhabitants/ km²	Independent	Density of inhabitants
Average Education	Discrete value	Independent	Average education level on scale of 0 to 4, where 4 indicated highest education like Masters or Ph.D. and 0 indicates no education
In house Service	Discrete value	Independent	Whether any in house service provided or not. It is 0 or 1.
Per Capita Income	PPP dollars	Independent	Per capita income of considered city or village in PPP dollars
Municipal Solid Waste	Kg	Dependent	Actual municipal solid waste in kg

The available data is imported using pandas data frame for the subsequent process. A sample of the same is shown below:

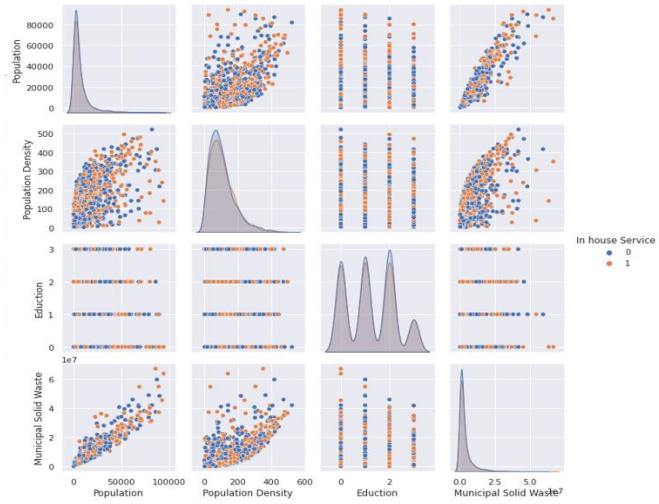

Figure 1: Pair plot of dependent variables categorized by in house service

It can be seen from Figure 1 that MSW is linearly dependent on population. Though it has some outliers, it is a safe assumption that it is linearly dependent [10]. In contrast, the relationship of MSW with population density is not linear. Hence logarithm of population density is taken.

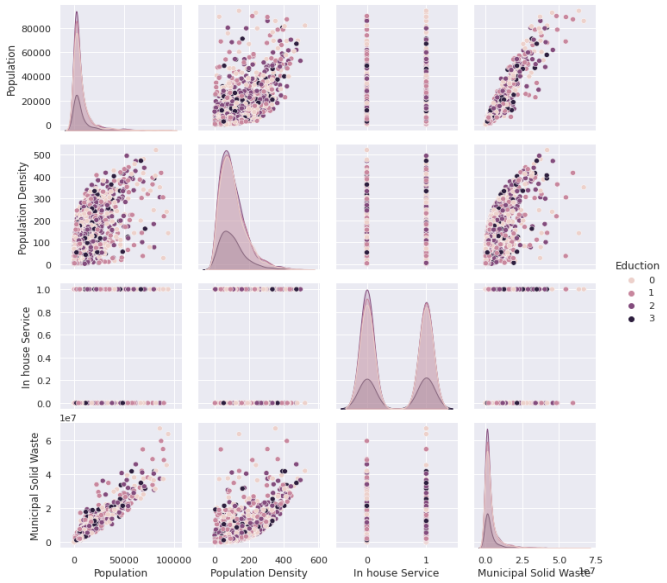

Figure 2: Pair plot of dependent variables categorized by education

The same pair plot is also visualized to find a relationship with the average education of the inhabitant as shown in fig. 2. Education of people also plays a vital role in MSW. Hence the same pair plot is also categorized by average education of region to gain insight for same. It was observed from fig 2 that the relationship between MSW and average education of the inhabitants was comparatively week.

3. RESULTS AND DISCUSSION

3.1 Statistical Analysis of Independent Variables

Primary exploratory data analysis is performed on the variables, and the box plot for the same is shown in fig. 3. From the box plot, it can be observed that outliers are there in some of the population and per capita income. From the exact figure, it can be seen that the population has a large no of outliers. However, from the data, it is verified that some cities have a large population compared to others; hence, this is not an anomaly but represents actual data. The same is valid for per capita income, as some cities have higher per capita income than others. Therefore, for this study, no data is dropped, considering that the outliers also pass accurate information. Population density was not found to have any outliers.

3.2 Linear Forecast Model

As education and in-house service are categorical variables, they have to be converted to a dummy variable to prepare data for linear regression. For this purpose, pd.det_dummies function is used. As dummies are created, a single education column is converted to Education_0, Education_1,..., Education_5. If all other columns have zero values, the first column will have an actual or one value. Hence it introduces multi-collinearity in itself. To avoid this, drop_first argument in pd.get_dummies is set as True. It will drop the first column, so the assumption of no multi-collinearity holds true.

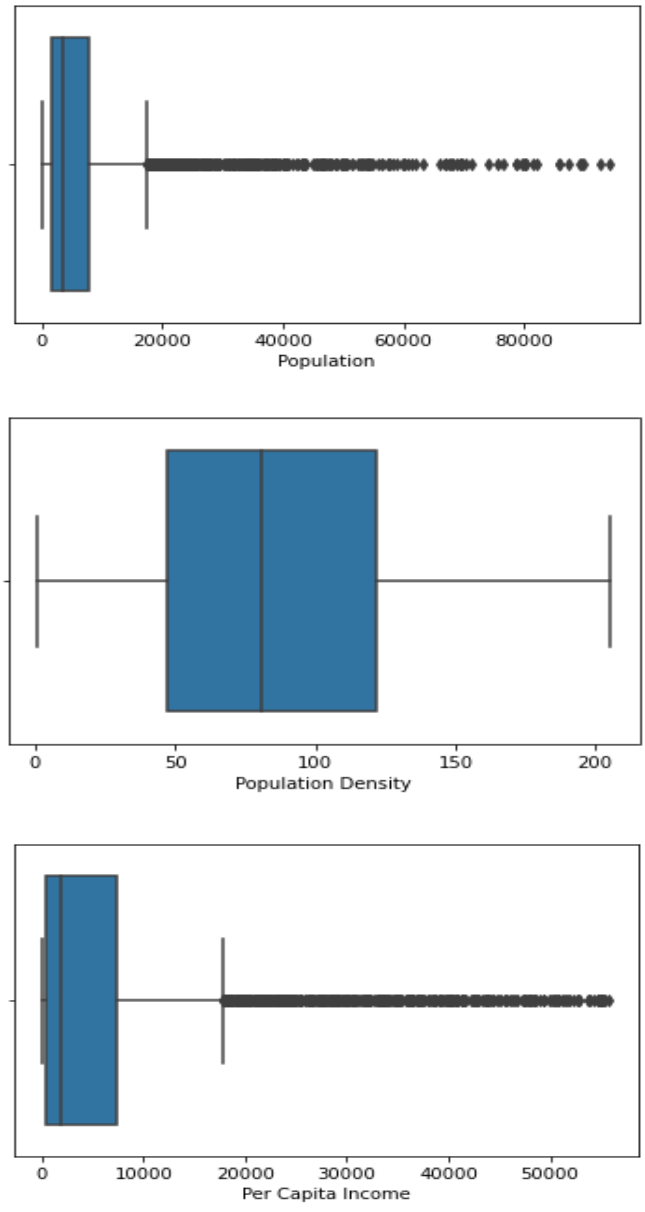

Figure 3: Box plot for continuous variables

	Population	Population Density	Per Capita Income	Municipal Solid Waste	Education_1	Education_2	Education_3	In house Service_1
0	22648	407.030647	102586	33956781	1	0	0	1
1	4952	135.269667	4904	2411867	1	0	0	1
2	3895	22.276110	3034	2159322	0	0	0	1
3	7140	62.739855	10195	3651915	0	0	0	1
4	12193	262.309573	29733	7195880	0	1	0	1
...	...	...	...	...	...	...	...	...
4292	5031	107.991900	5062	2113374	0	1	0	1
4293	2623	52.669641	1376	1240725	0	0	1	0
4294	13515	156.985971	36631	5554469	0	1	0	1
4295	4683	89.446720	4386	2172620	1	0	0	0
4296	5986	92.334159	7166	2102724	1	0	0	1

4297 rows × 8 columns

Figure 4: Final data frame for fitting model

As explained earlier, municipal solid waste is the dependent variable (y), and all other columns are independent variables.

In order to fit this into a linear model, Sci-kit learns library of python is used. The code used for the same is shown as follows,

```
[37] from sklearn.linear_model import LinearRegression
     lr = LinearRegression()
     X = df2.drop(columns=['Municipal Solid Waste'])
     y = df2['Municipal Solid Waste']
     lr.fit(X, y)

     LinearRegression()
```

As per the above image, a linear regression model is first imported. Then linear regression object is created for performing the regression task. As mentioned, the model is fitted with the independent and dependent variables.

Co-efficient and intercept of the model are found and shown below,

```
[34] lr.intercept_

     -425321.3057879135

[35] lr.coef_

     array([ 4.45637486e+02,  5.97846116e+03,  3.55935901e+00,  5.40007786e+04,
            -5.07457221e+02,  1.09006459e+04, -2.68332289e+04])
```

The score or R-squared value of the linear regression model is found for the authenticity and accuracy of the model. It is found as 0.918096, which is 91.81%. So, it is proven that the developed model is reliable for predicting municipal waste for municipalities.

```
[42] lr.score(X, y)

     0.9180961810704009
```

3.3 Evaluation of Linear Model

There are different evaluation matrices like MAE, MSE, and RMSE as explained in section 3.1. The present study considers three types of errors for evaluation and inferences.

From the developed model (lr), prediction can be found with the use of lr.predict() for known and unknown municipal solid waste, as shown below,

```
[46] y_pred = lr.predict(X)
     y_pred

     array([12493201.33297076,  2634802.62200822,  1427579.42265166, ...,
            6638690.10057273,  2265964.90816871,  2846954.78257005])
```

Here, the output array shows municipal solid waste generated in kgs considering different independent variables passed to it. Different errors for the calculated model are shown below:

```
[58] MAE = mean_absolute_error(y, y_pred)
     print('MAE : ', MAE)

     MAE :  15837.4758375192

[70] MSE = mean_squared_error(y, y_pred)
     print('MSE : ', MSE)

     MSE :  185913.0183025941

[74] RMSE = pow(mean_squared_error(y, y_pred), 1/2)
     print('RMSE : ', RMSE1)

     RMSE :  27431.17631927390
```

From the above errors, inferences can be made by considering MAE as 15,837. Using this method, it can be concluded that the absolute error in predicting municipality waste correctly is 15,837 kgs. So, while considering this method, it is advisable to consider this error while predicting MSW.

4. CONCLUSION

In the present research, multiple linear regression was used to develop a model to predict municipality solid waste for the municipalities of Italy. Different social and economic factors like population density, education, in-house service, and per capita income are considered. Out of all these parameters, the population factor was the most impactful factor for waste generation. Other parameters also affect the dependent variable, but have relatively less influence on the final results. Different linearity tests were also performed to check the model for linearity, multi-collinearity, homoscedasticity, and exogeneity.

Further, the model was developed to predict Municipal Solid Waste generated, given that users have a value of other independent variables. Different errors like MAE, MSA, and RMSE were also calculated to understand the data. It is advised to use this model for the value of range independent variable within the considered range. The present study uses limited data for the analysis due to its unavailability. The model is further improved by considering more data. Instead of using limited data, the model gives a relatively high accuracy of 91.81%.

REFERENCES

1. Edjabou, M.E., Jensen, M.B., Götze, R., Pivnenko, K., Petersen, C., Scheutz, C. and Astrup, T.F., 2015. Municipal solid waste composition: Sampling methodology, statistical analyses, and case study evaluation. Waste Management, 36, pp.12-23.

2. Nayak, P.C., Sudheer, K.P., Rangan, D.M. and Ramasastri, K.S., 2004. A neuro-fuzzy computing technique for modeling hydrological time series. Journal of Hydrology, 291(1-2), pp.52-66.

3. Beigl, P., Lebersorger, S. and Salhofer, S., 2008. Modelling municipal solid waste generation: A review. Waste management, 28(1), pp.200-214.

4. Pires, J.C.M., Martins, F.G., Sousa, S.I.V., Alvim-Ferraz, M.C. and Pereira, M.C., 2008. Selection and validation of parameters in multiple linear and principal component regressions. Environmental Modelling & Software, 23(1), pp.50-55.

5. Izquierdo-Horna, L., Kahhat, R. and Vázquez-Rowe, I., 2022. Reviewing the influence of sociocultural, environmental and economic variables to forecast municipal solid waste (MSW) generation. Sustainable Production and Consumption.

6. **Dataset:** Di Foggia, Giacomo (2022), "Municipal waste management cost and fee schemes", Mendeley Data, V1, doi: 10.17632/w5f9kg7743.1

7. Pany, P.K. and Ghoshal, S.P., 2015. Dynamic electricity price forecasting using local linear wavelet neural network. Neural Computing and Applications, 26(8), pp.2039-2047.

8. Estay-Ossandon, C., Mena-Nieto, A. and Harsch, N., 2018. Using a fuzzy TOPSIS-based scenario analysis to improve municipal solid waste planning and forecasting: a case study of Canary archipelago (1999–2030). Journal of cleaner production, 176, pp.1198-1212..

9. De Baets, S. and Harvey, N., 2020. Using judgment to select and adjust forecasts from statistical models. European Journal of Operational Research, 284(3), pp.882-895.

10. Azadi, S. and Karimi-Jashni, 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 management, 48, pp.14-23.

11. Ayeleru, O.O., Okonta, F.N. and Ntuli, F., 2018. Municipal solid waste generation and characterization in the City of Johannesburg: A pathway for the implementation of zero waste. Waste Management, 79, pp.87-97.