An evaluation of the APL and Independent House models

Samar Singh, PhD*

January 21, 2013

1. Introduction

The data for the APL apartments has been extensively reviewed for errors by the SD_SWM team, and recently became available last week after considerable checking.

The data have been analysed using the Open Source statistical language R, and the models shown are those that are appropriate and available in that resource. Some of the simpler models such as the Decision Tree model use very few parameters while the more complex models use a much larger set of parameters.

The models generated from the APL apartments data have been used to predict wet and dry waste values for Independent houses to explore the possibility that a common model will serve both purposes.

Seed values have been used to create a training set and a testing set for the APL apartments database. This means that though the observations are randomly chosen, they can be replicated by assigning the same seed value. Ideally one needs a large training set and a large testing set. Here I have used 90% of the observations for the training set and 10% for the testing set. Hence, using different seed values we have obtained multiple training sets, and hopefully defined a more robust means of measuring the validity of results.

2. Interpretations

- ▷ Wet waste means wet waste plus bio waste. Although we tried to collect bio-waste separately, it was discovered that many include bio-waste within their wet waste. As we do not normally open wet waste bags, unlike dry waste which is photographed, we chose to sum wet waste and bio waste across the board.
- ▷ Dry waste means everything other than Wet Waste. However, as we are photographing the contents of Dry Waste we have reasonable assurance of their appropriacy to this category.

Part I.
Apartments above poverty line

3. Approach

Exploratory work on the model showed that it was quite sensitive to the size of the dataset. For this reason we chose to adopt 5 different configurations of the **training set** using approx 90% of the

*TeamLead, SD_SWM Project

observations (319) in each set. The remaining observations (36) became the **test set**. This process was repeated 5 times so a total of 180 observations was used as the test set. A seed value is used on each occasion so that the set can be replicated later if needed.

A limitation of this approach is that as the training set varies each time the model in each case will be theoretically different. Another problem with this is that the test set is relatively small. This leads to reduced consistency in predictions.

The model also works on the idea of defining waste (wet or dry) per person per day. In practice the relationship between the number of persons and the waste generated is non-linear. However the data given is in the form of an average per person figure. The output is also given as that. However, the number of people in the primary models is the primary factor that explains the variance. Hence, the total waste generated should be able to be compiled by multiplying the number of people in a household by the waste generated per day.

4. Results - Wet waste generation

4.1. Predicted Values of Wet Waste Per Person (gm)

		Predicted Values			
Seed	Observed Wet Waste Per Person	Decision Tree Model	Random Forest Model	Linear Model	Neural Net Model
12.0	**265.1**	267.3	275.5	271.8	276.0
22.0	**311.3**	255.5	274.0	272.7	280.6
32.0	**239.8**	288.7	286.2	294.3	291.8
42.0	**291.8**	279.3	278.2	278.5	286.6
52.0	**283.2**	254.5	261.5	260.2	268.3

Figure 1: Predicted Values of Wet Waste Per Person (gm)

4.1.1. Explanation - Fig. 1

Figure 1 shows the predicted values for different seed values used to create a training set of 90% of the observations and a test set of 10% of the observations.

		%age error			
Seed	Observed Wet Waste Per Person	Decision Tree Model	Random Forest Model	Linear Model	Neural Net Model
12.0	**265.1**	−0.8	−3.9	−2.5	−4.1
22.0	**311.3**	17.9	12.0	12.4	9.8
32.0	**239.8**	−20.4	−19.4	−22.7	−21.7
42.0	**291.8**	4.3	4.6	4.5	1.8
52.0	**283.2**	10.1	7.6	8.1	5.3

Figure 2: Percentage error of predicted values of Wet Waste Per Person Per Day (gm)

4.1.2. Explanation - Fig. 2

Figure 2 shows the same observations as in Figure 1 but this time in terms of percentage error. We can see that in instances of seed values of 22 and 32 percentage error rises. For this reason, we aggregate the results of all test samples in the next section.

	Observed Wet Waste Per Person	Decision Tree Model	Random Forest Model	Linear Model	Neural Net Model
All obs.	**278.2**	269.1	275.2	275.6	280.7
Percentage Error		3.3	−2.2	−0.2	−1.8

Figure 3: Aggregated values of Wet Waste Per Person Per Day (gm)

4.1.3. Explanation - Fig. 3

This shows the results of aggregating all the test samples. This leads to a very small error of 2% or less for the Neural Net and the Linear models.

5. Dry waste generation

	Observed Dry Waste per person per day	Decision Tree Model	Random Forest Model	Linear Model	Neural Net Model
Test Instances(176)	58.7	61.8	62.7	64.9	65.2
Error in grams		−3.1	−4.0	−6.2	−6.5
StDev	36.5				

Figure 4: Predicted Values of Dry Waste Per Person Per Day (gm), Error in grams and St. Dev

5.1. Explanation - Fig. 4

Figure 4 shows the value of observed and predicted dry waste per person per day for the aggregated partitions of the database of 90% training and 10% testing. While the errors are large in percentage terms they are modest in terms of actual quantities and also when compared with the Standard Deviation of the observed values.

Part II.
Independent homes

6. Wet Waste generation

Currently, data is available for 63 independent houses. Of these 32(Set A) are homes in a gated colony and 31(Set B) are truly independent homes on a street. As the sample size is too small to develop a model, we have attempted to fit the Apartment model to these data. For the purposes of modeling, the entire dataset of apartments has been used to generate the model.

		Predicted Values from Apartment model			
	Observed	Decision Tree model	Random Forest Model	Linear model	Neural Net model
BothSets	251.4	283.5	272.0	251.7	245.0
SetA	289.1	291.4	280.5	268.4	258.0
SetB	212.4	275.3	262.9	234.0	231.0

Figure 5: Predicted Values of Wet Waste Per Person Per Day for Independent houses(gm)

6.1. Explanation - Fig 5

While the results of the total number of homes is very satisfactory in terms of the Linear and Neural Net models, the same cannot be said for the results of the Sets taken individually. This could be either due to the reduction in sample size or to the inappropriacy of the models to Independent homes. It may therefore be desirable to obtain a larger sample of independent homes - both gated and street based.

7. Dry waste generation

	Predicted values using Apartment model for Dry Waste				
Observed	Decision Tree model	Random Forest model	Linear model	Neural Net mode	Data
61.6	62.7	59.4	55.9	59.0	SetA+SetB
74.7	65.0	60.4	61.0	63.6	SetA
48.0	60.4	58.3	50.5	54.2	SetB

Figure 6: Predicted Values of Dry Waste Per Person Per Day for Independent houses(gm)

7.1. Explanation - Fig. 6

Here we are seeing for the total set i.e. A + B errors of only about 2.6 grams for the Neural Net model, these errors seems to be significantly larger in the case of Set A. It is possible that this is caused by the small size of the sample or it could be that Independent homes in gated colonies require a different model as the model seems to predict lower values in Set A and higher values than actual in Set B, albeit only marginally for the Linear and Neural Net models.

Figure 7 shows the distribution of predicted values which would lend itself to the view that the model is generally appropriate but needs to be tested with a larger sample. In a smaller sample the effect of outliers as seen in Fig 7 can be disproportionately large.

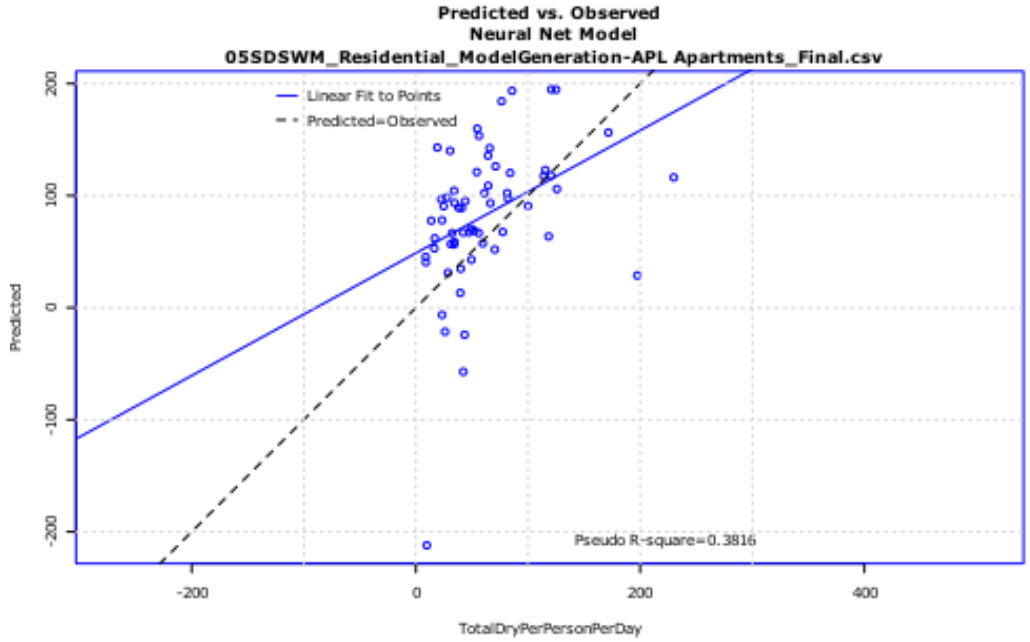

Figure 7: Predicted vs Observed Values of Dry Waste Per Person Per Day for Independent houses(gm)

8. Conclusions

The following conclusions are worthy of discussion:

- The apartments model (Linear/NN) works generally well for apartments but needs to be tested more extensively against independent houses.
- The total waste generated can be obtained by multiplying the predicted waste per person by the number of persons in the household.
- The models select widely varying number of parameters for calculation, with the Decision Tree model using only 7 and the Neural Network model using 20 parameters for wet waste. It is possible that the different models could be used for different requirements.

- We have to be open to the possibility that there are other more robust models which could be used.
- Before we make this data public the calculations should be replicated with the same data by another researcher to ensure there are no errors in intepretation or results obtained.
- On the assumption that wet and dry waste distributions should be close to normal distributions within the population as opposed to the sample, it seems likely that larger samples will provide better models.
- I will be discussing with the MIT team later this week, how we can use technology to get larger samples from crowd-sourcing.
- Fig. 8 shows the current distribution of wet and dry waste within the sample obtained from Apartments. We can see that the distribution of wet waste per person per day seems to be a bi-modal distribution. The dry waste distribution by contrast is unimodal though skewed to the right.

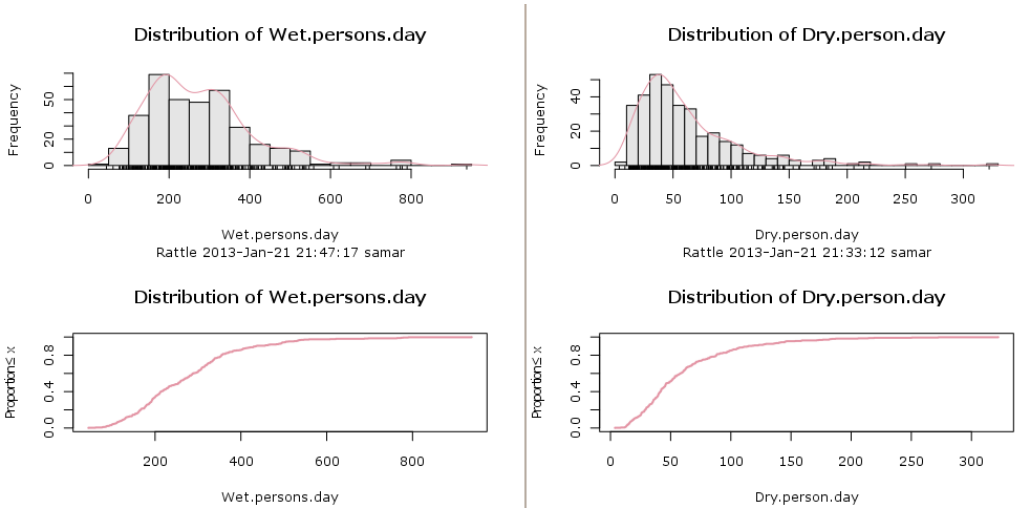

Figure 8: Distribution of wet and dry waste from surveyed apartments in grams.

Part III.
Acknowledgements

The modeling has turned out to be apparently more accurate than I had hoped for. However, none of this work would have been possible without the diligent efforts of the team that has been doing the groundwork - often difficult and frequently aggravating - for these many months. It has been a privilege to be on the receiving end of their considerable efforts. The team includes:

Manager, Research

- Akshay Yadav

Research Associates

- Sumiya Tarannum
- Krithi Venkat

Field Associates

- Melvin Deepak SK
- Tejaswi DK
- Praveen Kumar
- Ambarish BC