CIS 520 Spring 2018, Project Report

Predicting Road Width for United States MSA's: Towards a Measure of Road Land Usage in American Cities

Team Members:

Mian Umair Ahsan (PennKey: `ahsanm`; Email: `ahsanm@sas.upenn.edu`)
Ilan Gold (PennKey: `ilangold`; Email: `ilangold@sas.upenn.edu`)

Assigned Project Mentor:

TA-Saurav Bose

Team Member Contributions:

Team Member	Contributions
Ilan Gold	-Fit Baseline Models -Fit RMSE Models (Random Forest, MLPRegressor) -Data Collection
Mian Umair Ahsan	-Fit Keras Neural Network -Data Cleaning -Plotting and Statistical Analysis Subroutines

Code Submission:

[Submission on Canvas.]

Abstract

We seek to predict the widths of roads in Metropolitan Statistical Areas across the USA based on geographic features. Our loss functions were RMSE and MAPE. We found that while Random Forests provided the best RMSE, all RMSE based models under-performed severely on small roads (less than 6.2 meters). Thus we also fitted a Neural Network whose loss function was MAPE and saw a reduction of 5 percentage points overall, including large reduction on small roads. Furthermore, we hypothesize given evidence on measures of feature importance that binary features like MSA-location are good features - thus we hope that more data will allow finer binary features like Census Tract-level vectorizations.

1 Introduction

A statistic of interest in urban planning is how much land area of a Metropolitan Statistical Area (MSA) is taken up by roads. Road lengths are easily available simply by looking at a map - road widths are harder. Attempts have been made to use satellite imagery to classify roads as such but this can be computationally expensive and requires complete satellite imagery.

Our goal is to predict the width of roads based on a number of features coming from a combination of census data and from the hand-measured road-width data. That is, each road has been measured from its satellite image by hand, and associated with it is the location of the road (census tract as well as geo-coordinates) as well as its distance to the Central Business District (CBD) of the MSA. Our training/test data thus consists of the road widths along with their geographic information as well as attached census data, which includes features such as population and demographic information. We thus attempt to predict the width of a new road in some MSA given its features from the data set.

2 Related Work

We were not able to find any related work to this method of classifying i.e using non-image features to predict road width. However, the field of detecting roads from satellite imagery is robust but relies entirely on image-based data and is an open problem with many of its own difficulties. From a review of existing attempts at this problem, "The difficulties of road extraction from RS images lie in that the image characteristics of road features can be affected by the sensor type ... etc. In practice, a road network is too complex to be modeled using a general structural mode" [**Wang et al.**].

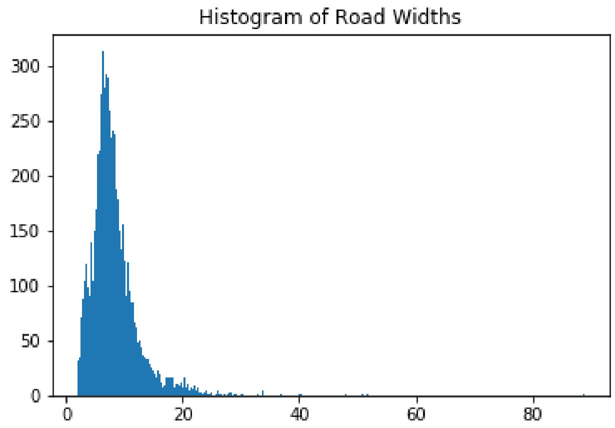
Furthermore, classifying all the satellite imagery of all the roads in all the cities of America would be a computationally expensive task to say the least. Thus, we seek to circumvent this issue and attempt a solution that relies on a more compact data source and could potentially be more efficient.

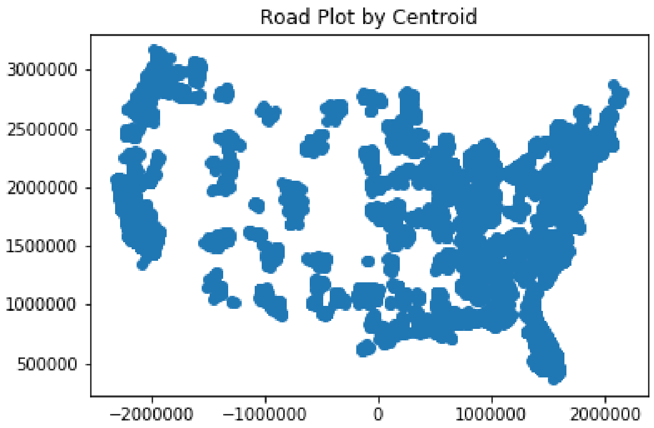
3 Data Set

Our datasets come from two places - a professor in the Institute of Urban Research for whom one of the authors works, and from the US Census. The census data is publicly available and was obtained using the R package tidycensus. We broke our data set up into a 70-30 split on 29098 observations for training and testing, respectively. We then standardized this split by always using random state of 42 so that we would get the same division every time.

First, we present our summary statistics for the width data, including a histogram (only from the training data set). Second, in order to in fact confirm that we have informative features in the Centroid measurements, we present a 2D plot of the road observations by their centroids - luckily it looks like the United States.

Road Width			
Mean	*Standard Deviation*	*Minimum*	*Maximum*
8.23	4.20	2.00	88.82





4 Problem Formulation

We have obtained data consisting of 29098 observations of road widths, their geographic location by census tract (as well as geo-coordinates, zip code, and MSA), and their distance from the CBD. To this we attached census data about population density and housing density. Using this as test/train data, we attempt to use learning algorithms to predict road-width based on features of the road. Our data also contained incorrect measurements as the owner of the data pointed out (i.e roads that are less than a meter or two wide, or roads that are over 100 meters wide) and these were either corrected using an auxilliary correction data set or deleted.

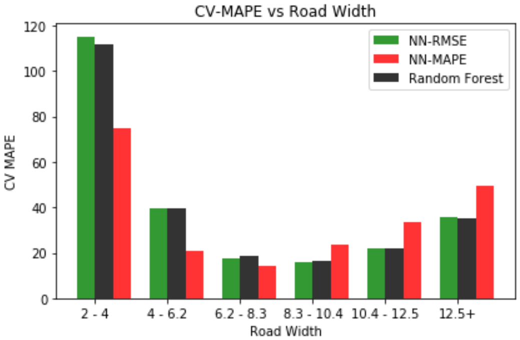
We decided to use both MAPE and RMSE to measure our performance. RMSE is a clear choice as it is a common function to use as an objective in regression and it is interpretable when compared to the Standard Deviation (i.e, how much more of the variance is explained by the model than by just predicting the mean). MAPE was less clear initially, but looking at a histogram of the data makes clear why we chose to use it - RMSE might fail to capture the difference between predicting

4 meters for a road of width 6 rather than predicting 50 meters for a road of width 52. We need a model that would be robust to handling such small length roads - if an unseen data set were to come along containing a lot of small roads, our model would not perform very well. For instance, a third of our data has road widths less than 6, so an RMSE $\simeq 4$ can produce poor estimates of road land usage in a crowded city where roads are significantly narrower.

5 Algorithms

We used least square regression as our baseline model to compare how much improvement other complex algorithms offered above the baseline. More complex models included Ridge regression, Random Forests, Boosted Decision Trees, and Neural Networks, with both RMSE and MAPE loss function. We expected Neural Networks with ReLU activation units and Decision Trees to be suitable models because our data contained many one-hot encodings. We used Keras library to implement Neural Network with MAPE loss and SciKit-Learn library for all other regression models in Python. We decided to implement Neural Networks with MAPE loss function because cross-validation MAPE on training data of models optimized with RMSE loss confirmed our suspicion that in order to minimize RMSE, these models were mispredicting on lowers width by a significant percentage. Figure 1 shows the MAPE occurred on various width ranges during cross-validation on our training data. Indeed, the results on test set in Figure 3 corroborate our observations from cross-validation.

Figure 1: MAPE on Cross-validation.



6 Experimental Design and Results

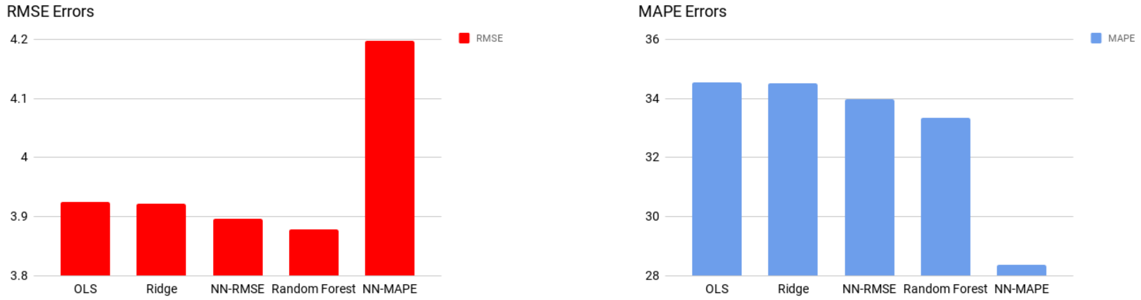
We created a $70-30$ training and test split of our data using SciKit by shuffling the data with a fixed random state. We decided to scale our data since we had both continuous and discrete variables (one-hot-vector encodings of Road Class and MSA) of various magnitudes and some of our machine learning algorithms were not scale invariant (e.g. Neural Network). We scaled our continuous features to zero mean and unit variance by standardizing them using the mean and standard deviation learned from the training data. We then fitted both the training and test features using these statistics.

Table 1: Parameters chosen from cross-validation.

Algorithm	*Parameters*
Ridge Regression	Regularization=1
Random Forest	Min Splits = 68
NN-RMSE	ReLU, 100 Layers, r=1
NN-MAPE	ReLU, 100 Layers, r=0

We carried out cross validation for RMSE to choose parameters for Ridge regression (regularizer), Random Forests (minimum number of instances in a node), and Neural Network-RMSE (number of hidden units, activation units and regularizer). For Neural Network-MAPE, we used cross validation for MAPE to select from among the same parameters as the other Neural Network. The best parameters found are summarized in Table 1.

Figure 2: Test Errors

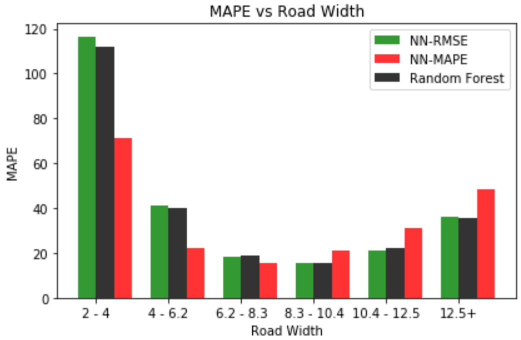


The bar plots show the RMSE and MAPE errors on the test set from model learned with parameters in Table 1. Best models w.r.t RMSE was Random Forest with a test error of 3.878 and the best model w.r.t MAPE was Neural Network-MAPE with test error of 28.372 percent.

We saw major improvement in mean percentage errors on narrower roads with NN-MAPE model, albeit, at the expense of percentage error on wider roads and overall RMSE. Figure 3. shows how percentage errors on various width ranges in our test set.

We report only Random Forests here as they outperformed other tree methods (see section 7).

Figure 3: MAPE on Test set.



7 Conclusion and Discussion

One issue with this project arises from the fact that we do not know a priori if our features are actually informative. Our evidence suggests that while they may not be excellent, they do in fact have some predictive power. Below we present an example figure in which we tack the RMSE and MAPE of the OLS baseline model as we add features in. While this is just a few ways of adding predictors (there are a combinatorial number of possibilities), the evidence is strong that our features do have predictive power, most strongly our binary features. Furthermore, Random Forests are often suggested as a method of feature selection as they offer a feature importance method based on the total reduction of the criterion (RMSE) brought by that feature over all trees in the Forest. Our top features are in Table 2. While this table may seem contradictory to our plots in Figure 4 (below), it is not as misleading as it might seem. Random Forests have been shown in experimental studies to be biased towards continuous features over binarized ones, as well as towards features that are not correlated to one another. [**Boulesteix et al**]. Given the importance of ACC Class 4 despite this, and the plots from the OLS baseline from before, we believe that more data would be beneficial in allowing us to improve our models as we would be able to use census-tract-level binarized predictors (as it stands, we only have about two observations per census tract).

Figure 4: Baseline Errors by Adding Predictors

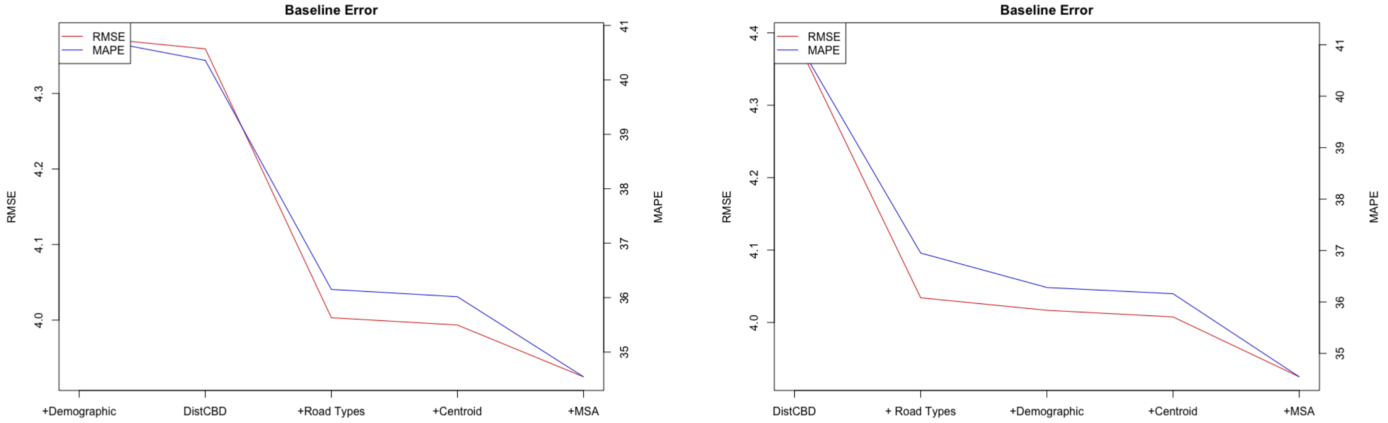


Table 2: Feature Importance.

1	ACC Class 4
2	Population Density
3	Centroid
4	Housing Ratio
5	Distance To CBD

It is also worth noting that our choice of Random Forest over Decision Trees as a final choice of tree-based regression was not unmotivated - Random Forest proved to be a better regressor than any boosted (or non-boosted) variation of decision trees. First, we cross-validated to obtain the parameters of the tree based on minimum RMSE (maximum depth and number of samples required for a node to split). Then, we attempted to boost the optimal tree. The best tree, with maximum depth of 5 and minimum-samples-to-split-on of 82, gave RMSE of 3.954 while the best boosting came from SciKit's AdaBoostRegressor (AdaBoostR2). The SciKit implementation of Regression boosting relies on a boosting criterion of absolute, square, or exponential loss. Given our previous discussion, though, we found ourselves interested in MAPE as well and therefore attempted a different boosting algorithm called AdaBoostRT, which relies on thresholding the MAPE as its boosting criterion. Neither version of boosted trees did as well as the Random Forest in RMSE. The AdaBoostR2 Trees performed only marginally better after cross validation, with RMSE 3.943 at .01 learning rate and exponential loss. The AdaBoostRT actually made the trees worse, with the best RMSE being 7.54 with a threshold of .9, which is remarkably high. It is worth noting that niether algorithm has a prover error bound, unlike the classification AdaBoost. [**Solomatine et al.**]

We were particularly pleased to find that using MAPE as a loss function in Neural Network allowed us to significantly improve prediction on small road widths as shown in Fig. 3, with an overall error of 28.4% compared to 33.4% for the next best model (Random Forest) w.r.t MAPE.

Acknowledgments

Prof. Shivani Agarwal, for her guidance in Office Hours.
Prof. Erick Guerra, for providing the data set and as well as domain knowledge.
Saurav Bose, for helping us think about the problem and our features.

8 Bibliography

Wu, Liang, and Yun-An Hu. A Survey of Automatic Road Extraction from Remote Sensing Images. Acta Automatica Sinica, vol. 36, no. 7, Mar. 2010, pp. 912922., doi:10.3724/sp.j.1004.2010.00912.
https://www.sciencedirect.com/science/article/pii/S2095756416301076

Boulesteix, Anne-Laure et al. Bias in Random Forest Variable Importance Measures: Illustrations, Sources and a Solution. SpringerLink, BioMed Central, 25 Jan. 2007
https://www.springer.com/article/10.1186/1471-2105-8-25

Solomatine, D.P., and Durga Shrestha. IEEE International Joint Conference on Neural Networks. AdaBoost.RT: A Boosting Algorithm for Regression Problems, Aug. 2004,
https://www.researchgate.net/publication/4116773_AdaBoostRT_A_boosting_algorithm_for_regression_problems