Impact on Total Market Value of Parcels: Introduction of the Street Car in Downtown Cincinnati

Exploratory Data Analysis, Predictive Analysis and Forecast

K. Rajesh Jagannath

Foundations of Data Science
Mentor: Anirban Ghosh
04/15/2017

Table of Contents

IMPACT ON TOTAL MARKET VALUE OF PARCELS: INTRODUCTION OF THE STREET CAR IN DOWNTOWN CINCINNATI 4
- INTRODUCTION 4
- OBJECTIVE 4
- MOTIVATION 4

PART I: DATA EXPLORATION OF THE ZONES UNDER STUDY 6
- DATA AND SOURCES 6
- EXTRACTION, TRANSFORMATION AND LOADING OF DATA 6
- EXPLORATORY DATA ANALYSES 7
- METHODOLOGY 8
- CONCLUSION (DATA EXPLORATION) 10

PART II: PREDICTION OF TOTAL MARKET VALUE OF PARCELS 11
- DATA AND SOURCES 11
- EXTRACTION, TRANSFORMATION, AND LOADING OF DATA 13
- FEATURE SELECTION 14
- DATA AND ANALYSES 16
- METHODOLOGY 19
- ETL METHODOLOGY 20
- STANDARDIZATION 20
- DIMENSION REDUCTION 20
- FORECAST MODEL DEVELOPMENT AND SELECTION 23
- PREDICTION 33
- VISUALIZATION OF RESULTS 34
- TABULATION OF FINDINGS 35

PART III: CONCLUSIONS AND FUTURE WORK 54
- CONCLUSIONS 54
- SUMMARY 59
- FUTURE WORK 59
- ACKNOWLEDGEMENTS 59
- PROBLEMS ENCOUNTERED DURING DEVELOPMENT OF METHODOLOGY 62

Impact on Total Market Value of Parcels: Introduction of the Street Car in Downtown Cincinnati

Introduction

The Cincinnati Streetcar is a modern streetcar system designed to link major employment centers in downtown and uptown, connecting through Cincinnati's historic Over-the-Rhine neighborhood.

It will operate 18 hours a day, 365 days a year.

Objective

The study's goal is to analyze and predict the "net effect" over the next 4 years on the economy of the City of Cincinnati within a buffer zone around the streetcar route by selecting Total Market Value of parcels from source data sets.

Motivation

Downtown is Cincinnati's largest employment center, with approximately 70,000 people working in the area everyday. It has been proven in cities from Atlanta to Seattle that fixed rails in the ground with thousands of potential riders draw new storefronts and businesses, as well as housing. These new businesses provide employment opportunity and boost a city's tax revenue. Furthermore, the **Annual Taxes** assessed on a parcel is a function of **Total Market Value** of the parcel.

During the construction phase, there may have been inconveniences to the neighborhood. Therefore, there are two camps of opinion -

- One opinion insists that the introduction of the streetcar is disruptive to the neighborhood (crowding, transient population, noise), and
- The other opinion is that it provides advantages such as, easy access to business, shops, dining and commuting to work and home and draws new business, expansion of storefronts, revenue from ridership, permit fees, property tax and restaurant license fee.

Three buffer zones around the streetcar route were established as shown below.

- CORE: The area shown in Red color is the CORE Buffer zone. The Streetcar runs through the center of this area along a North South corridor.
- CENTER: The area shown in Magenta color is the designated CENTER Buffer zone
- EDGE: The area shown in Green color is the EDGE Buffer zone

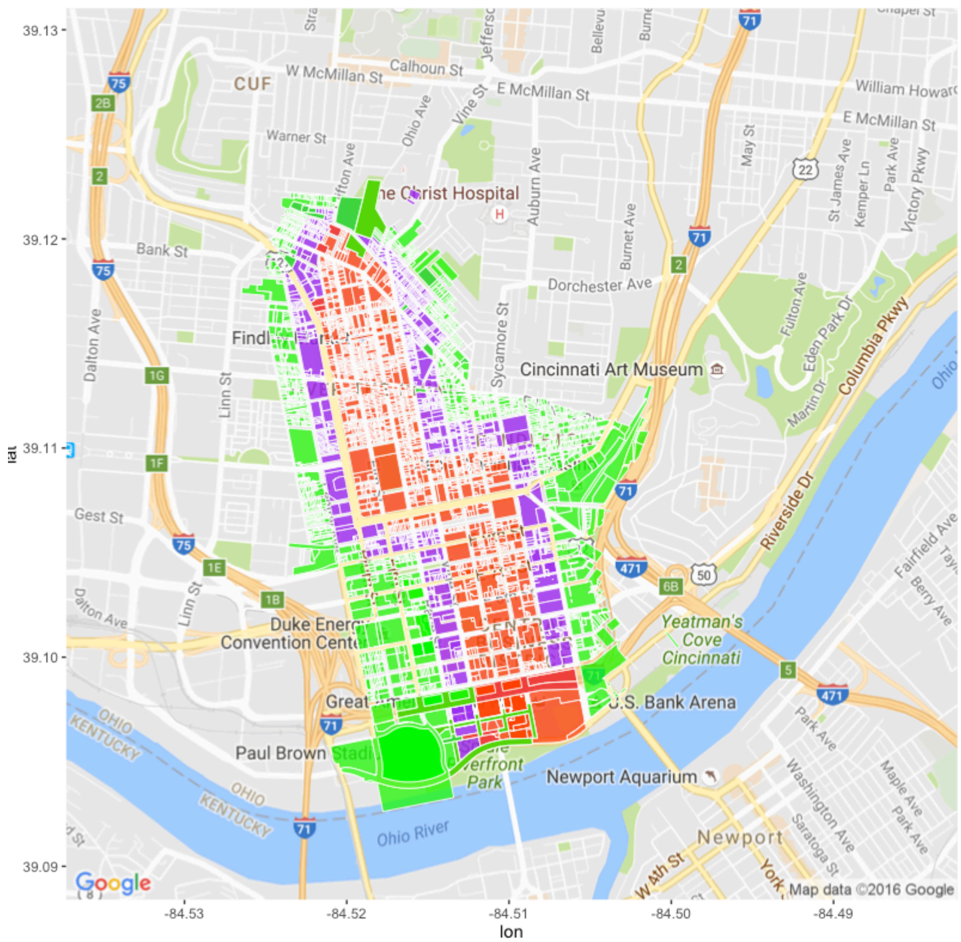


Figure 1 A ggplot of illustrating the buffer-zones CORE, CENTER and EDGE around the streetcar route

This report is organized in 3 parts**. Part I** covers descriptive analytics of the Buffer zone. **Part II** covers a forecast of economic impact due to the introduction of the streetcar system. **Part III** is a conclusion of findings.

Part I: Data Exploration of the Zones under study

Introduction: Part-I of this report covers the descriptive analytics of the zone under study. The sources of data, its confidentiality, and problems encountered with data, observations and conclusions are reported in this part.

Data and Sources

Data sources were Cincinnati Area Geographic Information Systems (CAGIS), City of Cincinnati, OH and the Hamilton County Auditor Office, Cincinnati, OH.
Confidentiality of data: The data is considered to be **Open Data** available to general public for use. To be sure, permission was obtained from CAGIS to be able to post the files on github for the purpose of this study

Buffer Area Parcels: There are three .csv files with an observation for each parcel in the three buffer zones under study.

a. StreetCarParcels_CORE.csv
b. StreetCarParcels_CENTER.csv
c. StreetCarParcels_EDGE.csv

Extraction, Transformation and Loading of Data

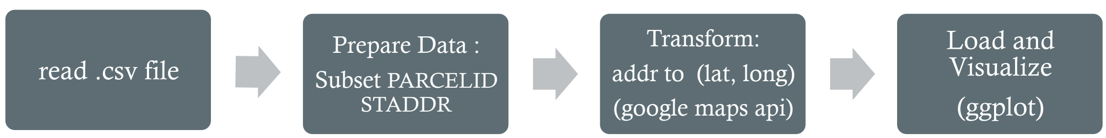
Buffer **zone under study**: The three buffer-zone parcel files were in .csv format. It was read in using read_csv. Features to obtain street address and parcel id were selected. The Street address was used to geocode the data to obtain longitude and latitude of the parcel.

Column name	Example Data	Description
PARCELID	7500010007	Unique id to identify parcels
EXLUCODE	C	Existing Land use Code e.g. Commercial
ADDRNO	1208	Address, street and type of street

Column name	Example Data	Description
ADDRST	SYCAMORE	
ADDRSF	ST	

Table 1: .csv files are used to identify the parcel id. of the three areas around the Street Car - Core, Center and Edge Buffer zones

The data was visualized for exploratory analysis. There are 900-1700 observations in each file. The file size is about 1.2 MB.



- StreetCarParcels_CORE.csv: n x p = 946 parcels x 67 columns
- StreetCarParcels_CENTER.csv: n x p = 1418 parcels x 67 variables
- StreetCarParcels_EDGE.csv: n x p = 1713 parcels x 67 variables

Exploratory Data Analyses

Buffer Zones under study: CENTER, CORE and EDGE

Scatter-plot of the location: Longitude, Latitude vs. Existing Land Use Code, visualizes the **expected 2-D distribution** of the parcels concentrated in the CORE, CENTER and EDGE zones in Figures 1, 2, and 3 below.

Figure 2: Scatterplot CORE Buffer Zone

Figure 3: Scatterplot CENTER Buffer Zone

Figure 4: Scatterplot EDGE Buffer Zone

Next, referring to Figures 4, 5, and 6 below, within the Buffer Zones, we find that the distribution with respect to Existing Land Use is **not uniform**. The distribution is skewed towards Multi-family, Mixed Used, Vacant, Commercial and Public/Semi-public parcels. Additionally, there are too many parcels classified as vacant lots, which was confirmed by looking at GIS (Geographical information systems) data.

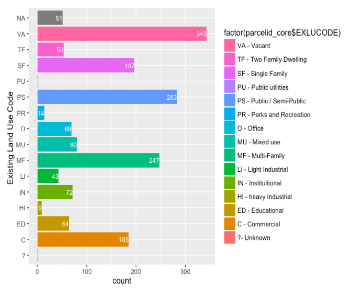


Figure 5. Histogram of parcels in CORE buffer zone.

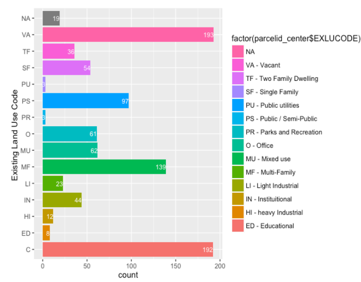


Figure 6. Histogram of parcels in CENTER buffer zone.

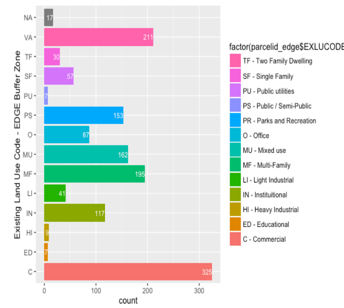


Figure 7. Histogram of parcels in the EDGE buffer zone

Methodology

By creating a 2-D Kernel Density plot, we find the areas of high Market Land Value to be centered on the Buffer Zone. There **is an unanticipated high-distribution in the center of the Downtown** in all the three plots. This is indicative of either **a problematic geocoding or the street addresses in the data are not correct**. In the scatterplot, Figure 2, Figure 3 and Figure 4, this problem is masked because the points are over-lapping each other in a single point in the center of the downtown. However, a 2-D Kernel Density Map reveals an unusually high concentration of observations in areas **not expected** to be in the CENTER and EDGE buffer zones.

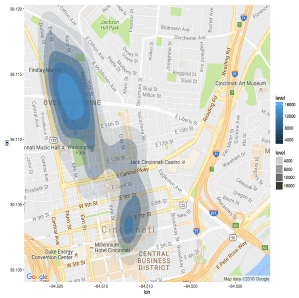


Figure 8: 2-D Kernel Density plot of CORE

Figure 9: 2-D Kernel Density plot of CENTER: High density of observations near Central Business District is not expected

Figure 10: 2D Kernel Density plots of EDGE: High density of observations near Central Business District is not expected

---

The data sets were analyzed further with staff at CAGIS, and a cleaner data set was obtained. In the clean data set, we observe the distribution of observations contained within the CORE, CENTER and EDGE buffer zones.

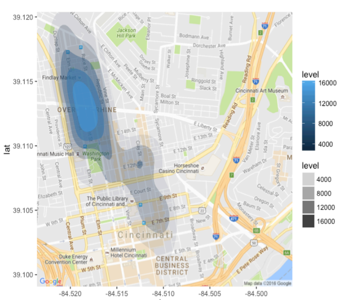


Figure 11: 2-D Kernel Density plot of the CORE parcels : AFTER – Clean Dataset provided by CAGIS

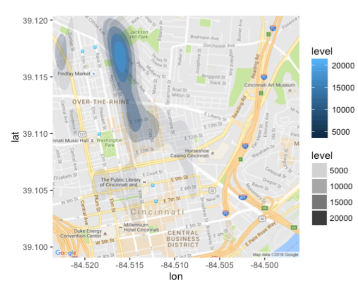


Figure 12: 2-D Kernel Density plot of the CENTER parcels: AFTER - Clean data provided by CAGIS - Central Business District are no longer there

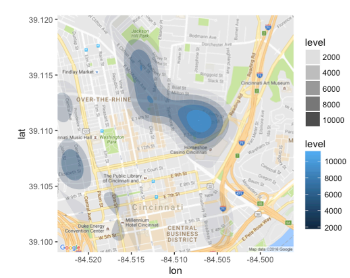


Figure 13: 2-D Kernel Density plots of the EDGE : After data was cleaned, observations are in line with expectations

Conclusion (Data Exploration)

- 3 .csv files, one each for CORE, CENTER and EDGE buffer zone were analyzed.
- When loaded, the data frame dimensions were 1713x8, 946x8 and 1418x8 respectively.
- Top 5 Land Use in each buffer zone were: Commercial, Multi-Family, Mixed Use, Semi-Public and Vacant
  - Examples of Vacant lots were Parking lots, which is to be expected near Commercial centers.
- Parcels with zero Annual Taxes value were analyzed further. They correspond to one of the several parcels owned by a single owner. The taxes are assessed on all the parcels and aggregated only on one parcel for billing convenience. The other parcels are therefore marked zero.
- Density map indicated that some geocoded co-ordinates are not spatially situated in the buffer zones as expected.
  - For example, in Figures 8 and 9, there is a high density of observations Near the Central Business District, which seem to be present in all 3 buffer zone.
  - Some of these observations do not have complete addresses for Google Maps Geocoding API to provide accurate longitude and latitude co-ordinates.
- Further analysis of the data set with the client, CAGIS, indicated that condominium parcels are also not correctly treated in the data set provided.
- Finally, a more accurate data-set was requested:
  - Figures 10, 11, and 12 illustrate a better distribution of the parcels in the expected buffer zones.
  - In particular, the high density of observations near the Central Business District in Figures 7, 8 and 9, prior to clean up is no longer observed.
  - This paves way for sub-setting data for Forecast Analysis (Part II).
  - Instead of using Google Maps geocoding, longitude and latitude co-ordinates were obtained from CAGIS directly.

Part II: Prediction of Total Market Value of Parcels

Data and Sources

The sources of data were Cincinnati Area Geographic Information Systems (CAGIS) and the Hamilton County Auditor Office, Cincinnati, OH.
Confidentiality of data: Data is considered to be **Open Data** available to general public for use. Data is real and not from any competition or academic study. Further, permission was obtained to be able to post the files on github for the purpose of this study.
Assessors Tax Information 2007-2015: The Hamilton County Auditor Office,Cincinnati, OH provided data for 9 years in Fixed Width Format in 9 files.

- taxinfo2007.txt
- taxinfo2008.txt
- taxinfo2009.txt
- taxinfo2010.txt
- taxinfo2011.txt
- taxinfo2012.txt
- taxinfo2013.txt
- taxinfo2014.txt
- taxinfo2015.txt

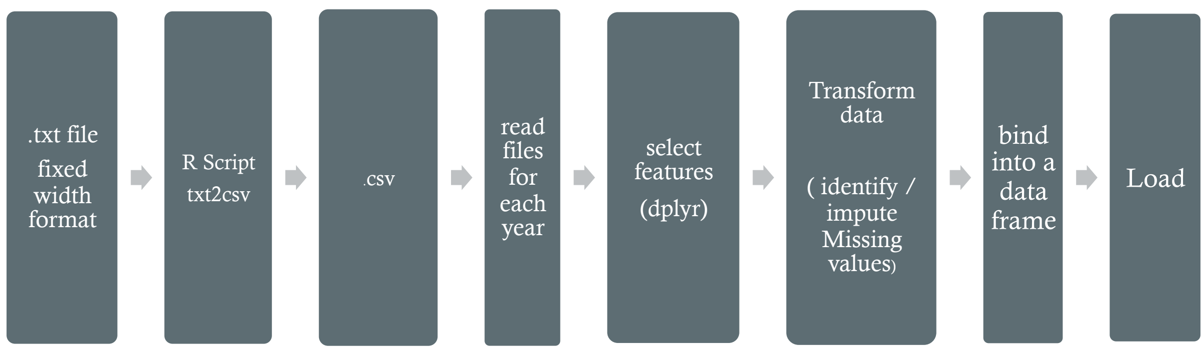
Column Name	Example Data	Description
PARCEL_ID	10001000100	Unique id for a parcel
LOC_STREET		Location e.g. for 2327 Sussex Ave., Cincinnati Oh
LOC_HOUSE_NO	2327	
LOC_ST_DESC	SUSSEX	
LOC_ST_IND	AV	
LOC_ST_DIR		

Column Name	Example Data	Description
VALID_SALE	Y	Yes or No
NUM_PARCEL	3	Number of Parcels
MKT_LAND_VAL	23000	Value of the Land
MKT_IMPR_VAL	140570	Market value of the Land
MKT_TOTAL	163570	Mkt. Total Val
ACRES	0.246	Acreage of the building
SALE_AMOUNT	116000	Sale Amount
SALE_DATE	20121129	Sale date in YYYYMMDD format
NEW CONSTR	N	Newly constructed building
ANNUAL_TAXES	3693.14	Annual Taxes Assessed
TAXES_PAID	3693.14	Annual Taxes Paid
DELQ_TAXES	6088.56	Delinquent taxes
FORECL_FLAG	Y	Tax Foreclosure Flag

Table 2: Features selected from Property Tax Information from years 2007 – 2015

Extraction, Transformation, and Loading of Data

Property Tax information 2007-2015: The original datasets were provided in fixed width format. An R script converted it to .csv file. The problem here was **each** of the groups of years 2007, 2008 and 2009-2014 and 2015 had different column widths. The field width was clearly documented. There are about 290,000 observations for each year. The final loaded data frame dimensions were 2,190,994 observations x 13



Feature Selection

From the tax-information of the data set a few features have been identified for selection. These selections are indicative of economic growth – Market Value, Assessed taxes, Revenue from Taxes paid, Sales data, Foreclosure Data and New Construction Flag

Sl	VARIABLE	Description
1	**PARCELID**	*A Unique identifier of the parcel*
2	**LOC_STREET**	*Address for plotting on ggplot or other package to identify spatial correlation*
3	**LOC_HOUSE_NO**	*Street Address Location + Latitude and Longitude*
4	**LOC_ST_DESC**	
5	**LOC_ST_IND**	
6	**LOC_ST_DIR**	
7	**cent_long**	
8	**cent_lat**	
9	**EXLU_CODE**	*Existing Land Use code*
10	**MKT_LAND_VAL**	*Market Value of land, Improvements, and Total*
12	**MKT_IMPR_VAL**	
13	**MKT_TOTAL**	
14	**ANNUAL_TAXES**	*Net Prop Tax revenue: Annual Taxes assessed, Taxes actually Paid, Delinquent Taxes and Tax Foreclosure*
15	**TAXES_PAID**	
16	**DELQ_TAXES**	

Sl	VARIABLE	Description
17	**FORECL_FLAG**	
18	**ACRE**	*Acreage to compute Property Value / sq. ft.*
19	**SALE_AMT**	*Sales data of Property: Amount, Sale Date, New Construction*
20	**VALID_SALE**	
21	**SALE_DATE**	
22	**NEW_CONSTR**	

Table 3: There are several features available in the data set for years 2007-2015. The features in the table above have been selected and are indicators of Market value of the parcel, Annual taxes, Acre-age, Sales Data. These are representative of the net economic effect.

PARCELID	2007	2008	2009	2010	2011	2012	2013	2014	2015
Over 290,000 observations/year									

Table 4: Parcel id uniquely identifies an observation. Each year from 2007-2015 has a subset of features shown in Table 10. This data is not tidy data and will need to be transformed using tidyr into Table 12.

During the course of the investigation, it was decided to limit the scope of the study to MKT_TOTAL_VAL, which stands for Total Market Value. It is the assessed value of a parcel including the land and the improvements (buildings, structures on the land). This decision to limit the scope came about because of time, resources and the effort required to clean the data.

YEAR	PARCEL_ID	MKT_TOTAL_VAL
2007	0010001000100	160,000
	…	
	…	
2007	…	
2008	0010001000100	160,000
	…	
2008		
2015	0010001000100	163,570
2015		

Table 5: The final data from df_MKT_TOTAL_VAL 2,190,994 observations of 3 variables – YEAR, PARCEL_ID and MKT_TOTAL_VAL

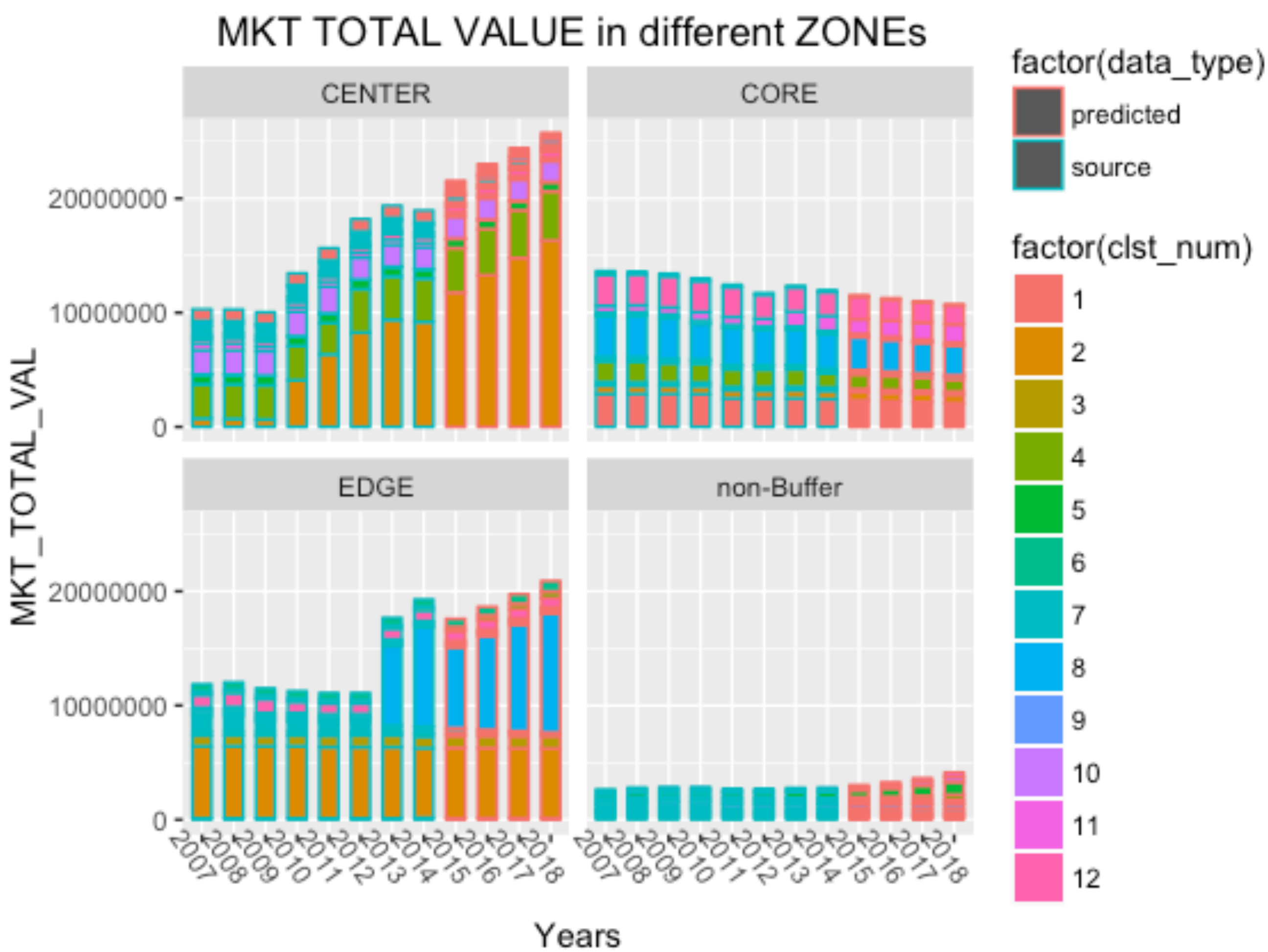


Figure 14: comparison of MKT_TOTAL_VAL in dollars in different zones. For comparison, 1224 parcels were chosen randomly from over 290,000 parcels in the Hamilton County. These were compared against CORE, CENTER, and EDGE buffer zones. Source data for years 2007-2014 (histograms outlined in teal color) along with forecast data for years 2015-2018 (histograms outlined in red) is displayed in the histogram.

- Parcels in Buffer zones have had large MKT_TOTAL_VAL compared to the non-Buffer Zone.
- The CENTER zones exhibit largest increase in MKT_TOTAL_VAL.
- Then it is closely followed by the EDGE zone.

- The changes in the CORE zone is the least.
- It is possible that the demand on properties situated in the CENTER and EDGE zones, is dictated by just-the-right-proximity to the streetcar. Too close seems less desirable.

CENTER ZONE vs. EDGE ZONE

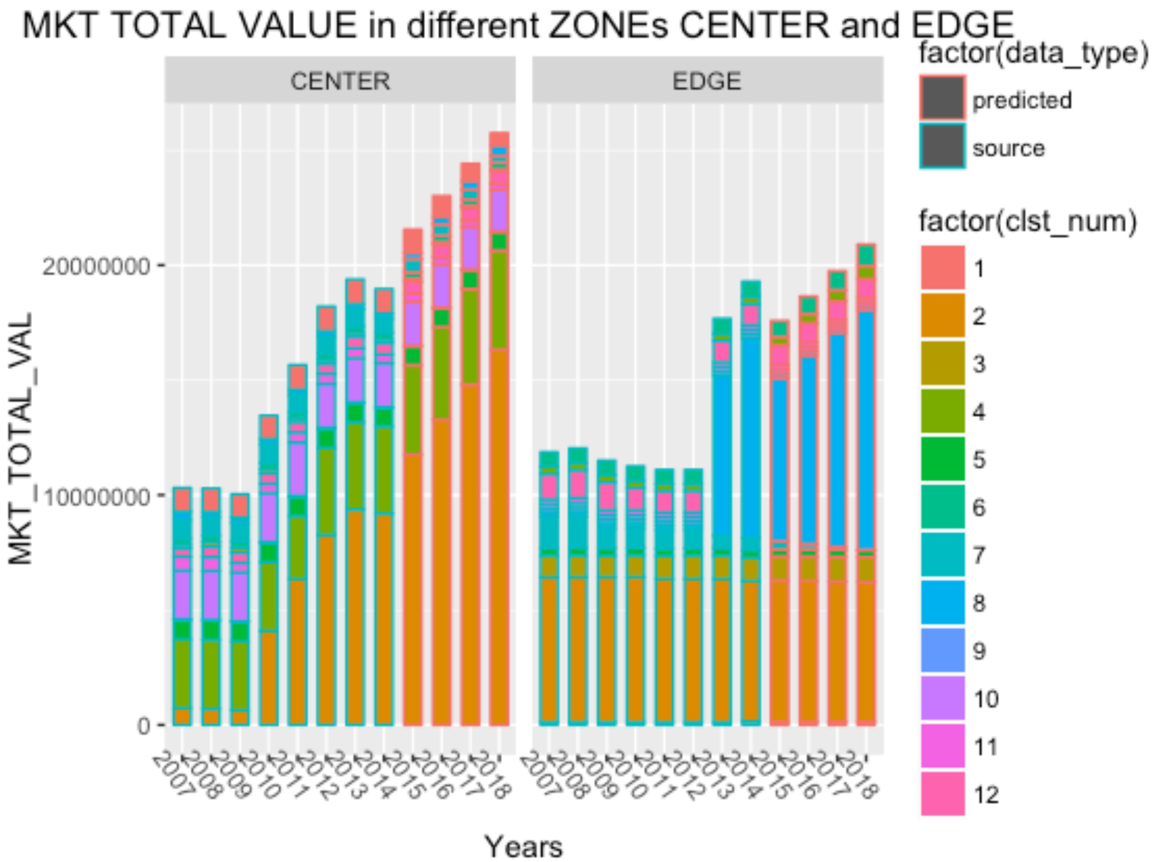


Figure 15:Comparison of CENTER and EDGE zones

A comparison between CENTER and EDGE zones, illustrated in the histogram in Figure 15, (above) shows, that the forecast for MKT_TOTAL_VAL for the years 2015-2018 of the CENTER zone is more than the EDGE zones for the years 2016-2018. Within the CENTER zone, the properties within cluster 2 show most significant growth in MKT_TOTAL_VAL.

CORE ZONE vs. NON-BUFFER ZONE

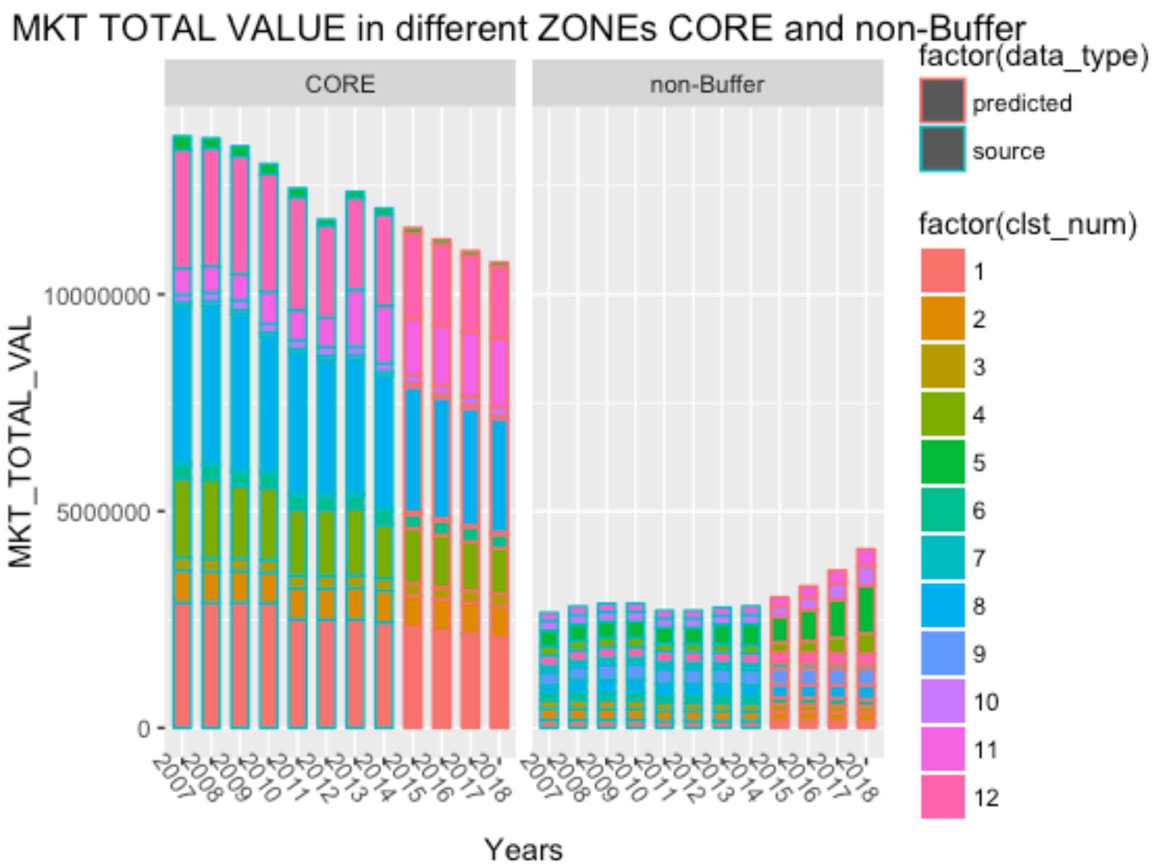
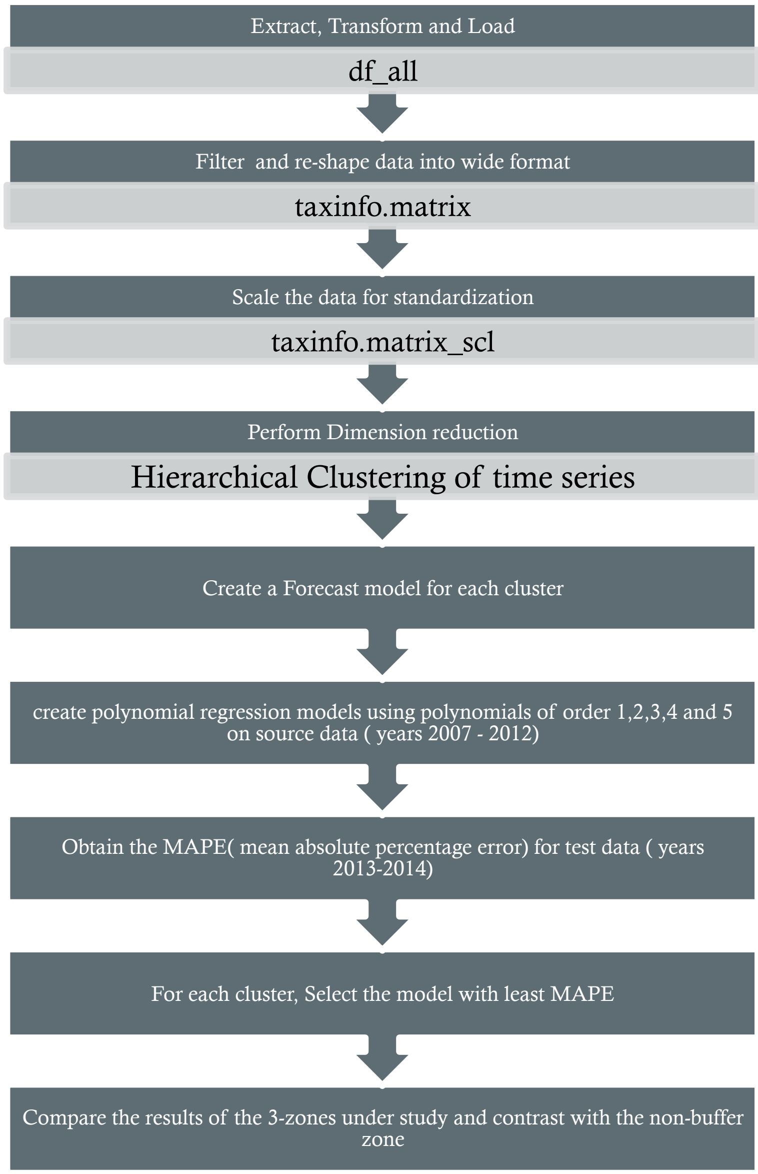


Figure 16: Comparison of CORE and non-Buffer zones

The non-Buffer zone trend shows an overall positive MKT_TOTAL_VAL trend even though it showed periods of declining trends in some of the years viz.,2010-2013. The prediction models a forecast of increasing values for the years 2015-2018. The CORE buffer zone model predicts a decline in MKT_TOTAL_VAL. Within the CORE zone, parcels in Cluster 1 show the greatest declining trend.

Methodology

The following flow chart illustrates an overview of the methodology



ETL Methodology

Step 1: Nine data frames, representing 9 years (2007-2015) of data were bound into a single data frame df_all. The original data included 13 predictors X 2,191,115 observations.

Step 2: To scope down the project, it was decided to focus on MKT_TOTAL_VAL predictor. The annual tax collected is a function of the MKT_TOTAL_VAL. It represents the land value and the value of the improvements (i.e. buildings on the land).

Step 3: To create a time series object, the data-frame df_all, had to be **re-shaped** into a wide format, entries with NA removed, and duplicate entries removed. Further, only data with 8 complete years (2007-2014) of data was used for study (detailed explanations in the Problems section). The final data frame was a matrix of 8 observations X 241748 predictors. There were only 8 observations / parcel id as the assessment is done on an annual basis.

Standardization

Step4: A **time series object** and analysis for every one of 241,748 parcel-ids **each with a frequency of 1 year** was not practical. So, as a next step it was decided to use a learning sample with a size of about 1224 parcels picked at random from a population of 241,748 parcels. 1224 was the mean sample size of the 3 buffer zones under study. This represents about 0.51699 % of the population size.

Using the **`scale()`** function, the MKT_TOTAL_VAL for each parcel-id was **standardized** (number of standard deviations away from the mean) in preparation for clustering of variables.

Dimension Reduction

Step 5: Clustering of variables is useful for **dimension reduction**.

Clustering of variables is as a way to arrange variables into homogeneous clusters i.e. groups of variables which are strongly related to each other and thus bring the same information. In this study, parcels with similar MKT_TOTAL_VAL trends over 8 years, were clustered together. For each zone under study, using **`ClustOfVar library`**, an ascendant hierarchical clustering of **PARCEL_IDs** was performed using **`hclustvar ()`** function. The aggregation criterion is the decrease in homogeneity for the cluster being merged. The homogeneity of a cluster is the sum of the squared correlation between the variables and the center of the cluster, which is the first principal component of PCAmix.

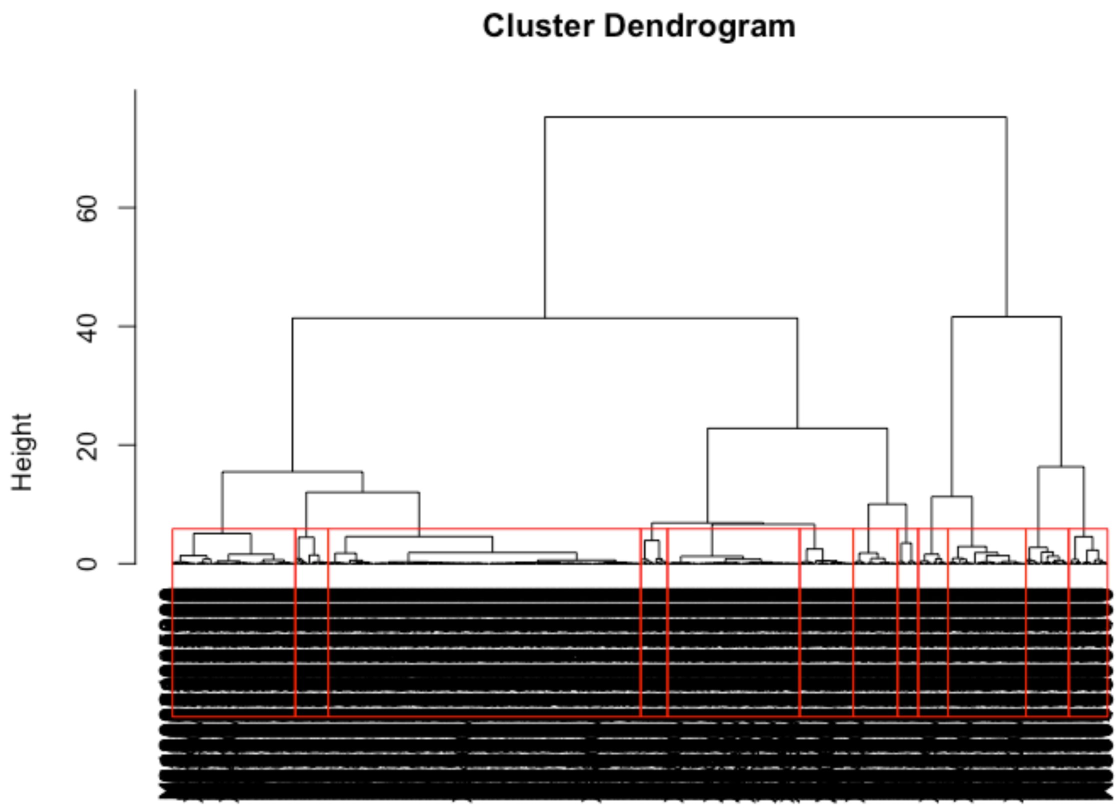


Figure 17: Dendogram of non-Buffer zone of 1224 sample parcel ids from the Hamilton County

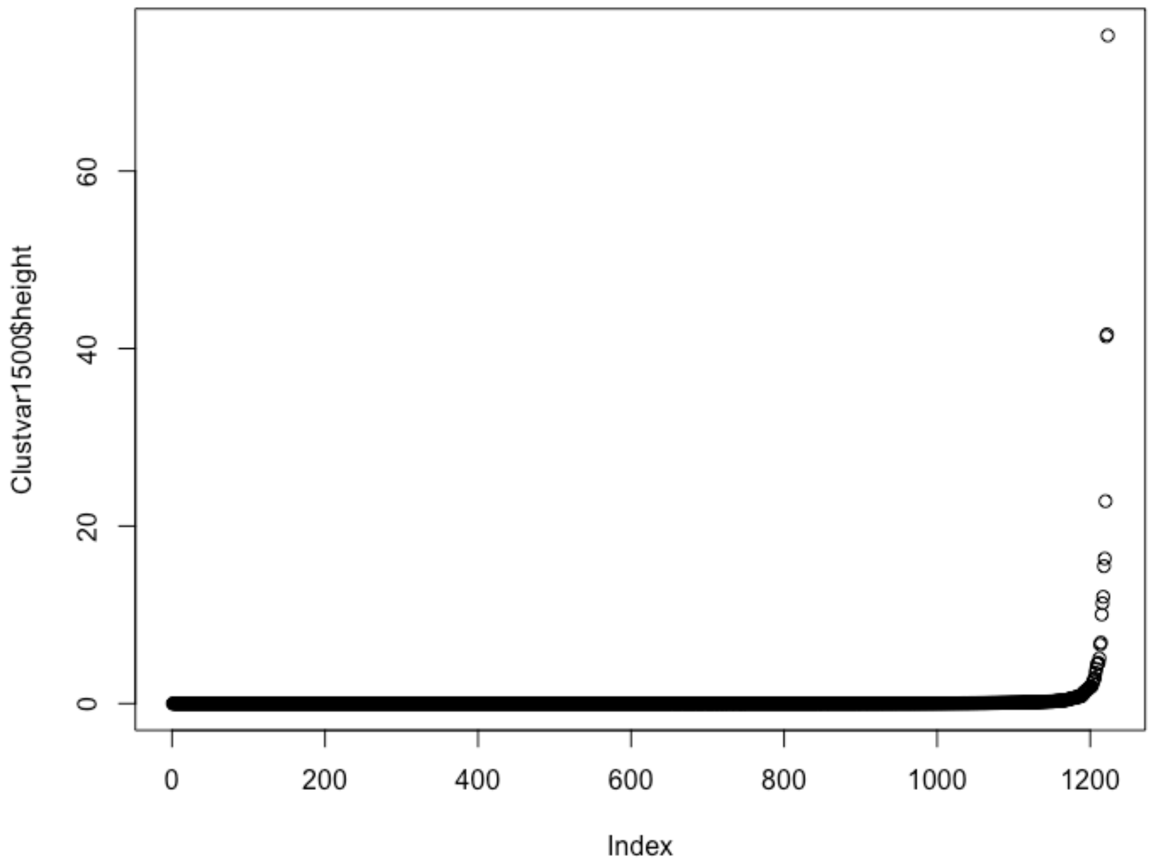


Figure 18: Based on intuition and the plot of height variable of the hclust object, a cluster of 12 seemed reasonable choice for further analyses

Based on the plot of the height variable of hclust variable as well as intuitions, cluster size of k = 12 was chosen.

Step 6: Polynomial regression

Polynomial regression extends the linear model by adding extra predictors obtained by raising the predictors to a power. This approach provides a simple way to provide a non- linear fit to data. For each of the clusters, cluster 1 through cluster 12, a polynomial regression model of the order - 1(linear), 2(quadratic), 3(cubic), 4 and 5 was created. The **training set** for each cluster was created for each polynomial order using the years 2007-2012. The **test set** was 2013-2014.

Step 7: Model Selection:

For each cluster and for each polynomial order, the Mean Absolute Percentage Error (MAPE) was computed for the test set. Variable **`y`** is actual value, and the variable **`yhat`** is the predicted value of the model.

Step 8: Visualize the Mean Absolute Percentage Error (MAPE): For each zone, the MAPE was calculated and visualized on a plot. With a user-defined function call, the actual vs. predicted MKT_TOTAL_VAL for each cluster and for each polynomial order was plotted. This plot reveals, that models with higher order polynomials have higher error on the test-data. The linear and quadratic order of polynomials is sufficient to describe most clusters, and at the most a cubic equation.

EDGE MAPE Visualization and Tabulation

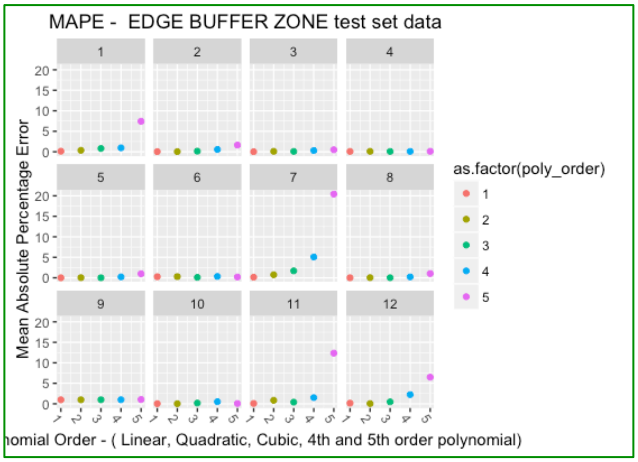


Figure 19. Mean Absolute Percentage error for EDGE buffer zone . Each panel in the facet represents the MAPE for that cluster. X- axis is order of polynomial. The Y-axis is the Mean absolute Percentage error. It is observed, that higher the polynomial order of the model, higher the error. As an example, for cluster 10, a polynomial of order 2 has the minimum MAPE and the predictive model od order 2 is chosen from out of 5 models.

ZONE	mape	poly_order	cluster
EDGE	**0.130266228**	**1**	**1**
EDGE	0.352409434	2	1
EDGE	0.805962756	3	1
EDGE	0.961419082	4	1
EDGE	7.430035127	5	1
EDGE	0.03599647	1	2
EDGE	0.042011492	2	2
EDGE	0.144317939	3	2
EDGE	0.562098914	4	2
EDGE	1.63911057	5	2
EDGE	0.016651428	1	3
EDGE	0.083331485	3	3
EDGE	0.104229433	2	3
EDGE	0.302099441	4	3
EDGE	0.489966503	5	3
EDGE	0.073921931	3	4

ZONE	mape	poly_order	cluster
EDGE	0.079867675	4	4
EDGE	0.082291995	1	4
EDGE	0.101789255	2	4
EDGE	0.105465291	5	4
EDGE	0.003645521	1	5
EDGE	0.009309043	3	5
EDGE	0.033841576	2	5
EDGE	0.223668396	4	5
EDGE	0.983326975	5	5
EDGE	0.102201592	3	6
EDGE	0.18812537	5	6
EDGE	0.260604994	1	6
EDGE	0.285571349	2	6
EDGE	0.332964935	4	6
EDGE	0.142685244	1	7
EDGE	0.742903157	2	7
EDGE	1.698010483	3	7
EDGE	5.072946484	4	7
EDGE	20.41624547	5	7
EDGE	0.008089368	1	8
EDGE	0.019598125	3	8
EDGE	0.03729133	2	8
EDGE	0.208665237	4	8
EDGE	1.025348029	5	8
EDGE	0.976711297	1	9
EDGE	0.979081715	2	9
EDGE	0.983982979	3	9
EDGE	0.997468716	4	9
EDGE	1.047297127	5	9
EDGE	0.032591925	1	10
EDGE	0.032988102	2	10
EDGE	0.065901839	5	10
EDGE	0.179803786	3	10
EDGE	0.514891729	4	10
EDGE	0.082420641	1	11
EDGE	0.402413533	3	11

ZONE	mape	poly_order	cluster
EDGE	0.854559938	2	11
EDGE	1.512671045	4	11
EDGE	12.35432366	5	11
EDGE	0.055122975	2	12
EDGE	0.153463021	1	12
EDGE	0.47816778	3	12
EDGE	2.232913354	4	12
EDGE	6.487499005	5	12

Table 6: The Table above depicts the MAPE for different clusters in the EDGE buffer zone. All the rows marked in blue font, represent the best model out of 5 polynomial order. As an example, For Cluster 12, polynomial order of 2 is selected as the its MAPE – 0.055122975 is the least amongst all the models.

CENTER ZONE MAPE

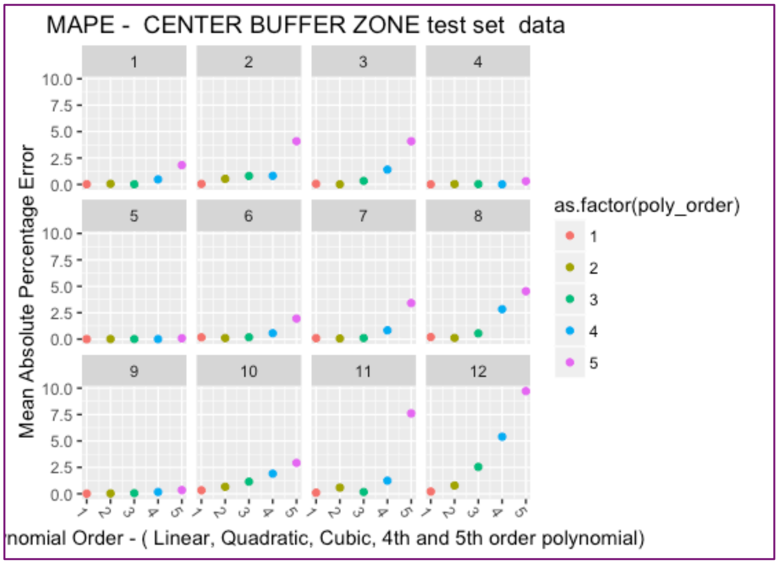


Figure 20 : For the CENTER buffer zone, the mean absolute percentage error was plotted against the polynomial order of the prediction model. Again, the higher order models don't provide and advantage. A polynomial order of 3 provides a predictive model with MAPE below 3%. As an example, for cluster 2, a linear model has the minimum MAPE and the predictive model of order 1 is chosen from out of 5 models.

ZONE	mape	poly_order	cluster
CENTER	0.016343028	3	1
CENTER	0.021384836	1	1
CENTER	0.056176972	2	1
CENTER	0.477860303	4	1
CENTER	1.827638765	5	1
CENTER	0.048927778	1	2
CENTER	0.527791297	2	2
CENTER	0.806746254	3	2
CENTER	0.815373041	4	2
CENTER	4.089342697	5	2
CENTER	0.010773457	2	3
CENTER	0.056878019	1	3
CENTER	0.3333754	3	3
CENTER	1.405540418	4	3
CENTER	4.086352471	5	3
CENTER	0.00843079	1	4

ZONE	mape	poly_order	cluster
CENTER	0.0100403	4	4
CENTER	0.029542063	2	4
CENTER	0.03338133	3	4
CENTER	0.305243261	5	4
CENTER	0.001235563	1	5
CENTER	0.004612875	4	5
CENTER	0.008211826	3	5
CENTER	0.009270585	2	5
CENTER	0.081217571	5	5
CENTER	0.10045817	2	6
CENTER	0.167168318	1	6
CENTER	0.17211478	3	6
CENTER	0.555049657	4	6
CENTER	1.941119623	5	6
CENTER	0.054077775	2	7
CENTER	0.092163761	1	7
CENTER	0.099078205	3	7
CENTER	0.832376163	4	7
CENTER	3.410496159	5	7
CENTER	0.111232247	2	8
CENTER	0.193387079	1	8
CENTER	0.545576374	3	8
CENTER	2.831905561	4	8
CENTER	4.532482291	5	8
CENTER	0.01659657	1	9
CENTER	0.031085984	2	9
CENTER	0.058165354	3	9
CENTER	0.172687635	4	9
CENTER	0.345927555	5	9
CENTER	0.329115079	1	10
CENTER	0.665319571	2	10
CENTER	1.147097496	3	10
CENTER	1.903876286	4	10
CENTER	2.92404141	5	10
CENTER	0.101459588	1	11
CENTER	0.168604557	3	11

ZONE	mape	poly_order	cluster
CENTER	0.584824152	2	11
CENTER	1.242655976	4	11
CENTER	7.607897508	5	11
CENTER	0.209887178	1	12
CENTER	0.78164799	2	12
CENTER	2.544333123	3	12
CENTER	5.398389057	4	12
CENTER	9.719800356	5	12

Table 7: The Table above depicts the MAPE for different clusters in the CENTER buffer zone. As an example, For Cluster 11, polynomial order of 1 is selected as its MAPE – 0.0101459588 is the least amongst all the models

CORE ZONE MAPE

Visualization

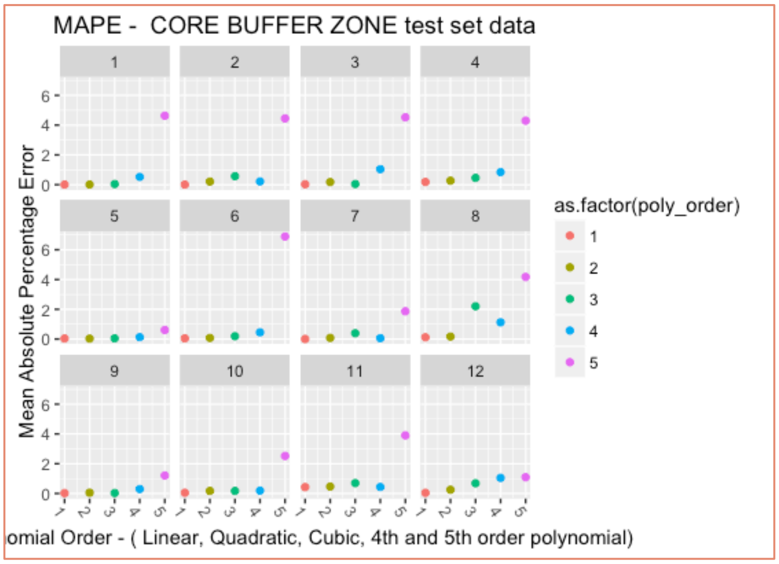


Figure 21 : For the CORE buffer zone, the mean absolute percentage error was plotted against the polynomial order of the prediction model. Again, the higher-order models don't provide any advantage. A polynomial order of 3 provides a predictive model with MAPE below 3%. As an example, for cluster 2, a polynomial of order 4 has the minimum MAPE and the predictive model of order 4 is chosen from out of 5 models

Tabulation

ZONE	mape	poly_order	cluster
CORE	0.004872816	1	1
CORE	0.008516562	2	1
CORE	0.039673932	3	1
CORE	0.522016619	4	1
CORE	4.632403954	5	1
CORE	0.005290909	1	2
CORE	0.206713115	2	2
CORE	0.216460808	4	2
CORE	0.567874555	3	2
CORE	4.450788902	5	2
CORE	0.029460522	1	3
CORE	0.04174543	3	3
CORE	0.174133769	2	3
CORE	1.03867621	4	3

ZONE	mape	poly_order	cluster
CORE	4.524604578	5	3
CORE	0.181560082	1	4
CORE	0.269052816	2	4
CORE	0.457535536	3	4
CORE	0.842600115	4	4
CORE	4.297242856	5	4
CORE	0.028808304	2	5
CORE	0.037541207	1	5
CORE	0.047344398	3	5
CORE	0.130659526	4	5
CORE	0.609662795	5	5
CORE	0.04827739	1	6
CORE	0.075161022	2	6
CORE	0.192299896	3	6
CORE	0.454564776	4	6
CORE	6.89160822	5	6
CORE	0.005970698	1	7
CORE	0.058469101	4	7
CORE	0.079588256	2	7
CORE	0.391709144	3	7
CORE	1.869740956	5	7
CORE	0.120563564	1	8
CORE	0.166185308	2	8
CORE	1.124275032	4	8
CORE	2.203831865	3	8
CORE	4.172720001	5	8
CORE	0.02362266	1	9
CORE	0.026249237	3	9
CORE	0.050050992	2	9
CORE	0.294979312	4	9
CORE	1.207071956	5	9
CORE	0.047302574	1	10
CORE	0.169080039	3	10
CORE	0.172359721	2	10
CORE	0.191077965	4	10
CORE	2.519043797	5	10

ZONE	mape	poly_order	cluster
CORE	0.429630049	1	11
CORE	0.439375707	4	11
CORE	0.465385367	2	11
CORE	0.697725879	3	11
CORE	3.902736665	5	11
CORE	0.041833768	1	12
CORE	0.252656675	2	12
CORE	0.686313095	3	12
CORE	1.047413735	4	12
CORE	1.090634266	5	12

Table 8:. The table above depicts the MAPE for different clusters in the CORE buffer zone. As an example, For Cluster 12, polynomial order of 1 is selected as its MAPE – 0.041833768 is the least amongst all the models.

Prediction

Step 8: Next, we call a user-defined function non_linear_regression_model (). The pseudo-code is as follows:

- For each buffer zone in the list {CORE, CENTER and EDGE}
  - For each polynomial order in the list {linear, quadratic, cubic, 4^{th} order and 5^{th} order}
    - For each Cluster in a buffer zone
      - Create a time series object on the row means (years 2007-2014) and create a non-linear regression model using
      - `df_mkt_val3 <-lm (ts_train_cl_num [1:8] ~ poly(Time,poly_order))`
      - Return a data frame by binding the source MKT_TOTAL_VAL for years 2007-2014 and Predicted MKT_TOTAL_VAL for years 2015-2018
      - Visualize the plot

Visualization of Results

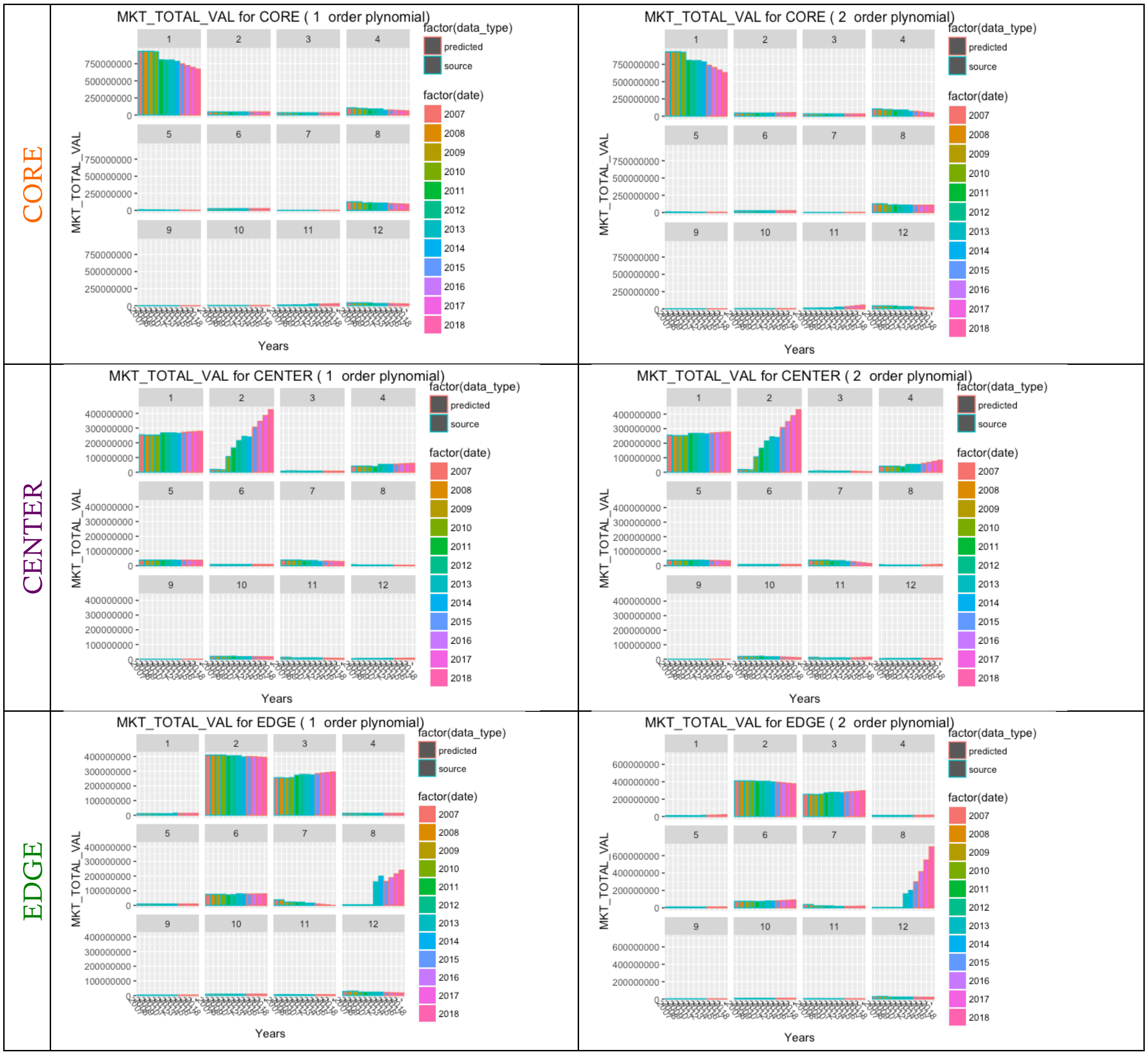


Figure 22: Visualization of MKT_TOTAL_VALUE of the three buffers. X axis in Years (2007-2014 – source years , 2015-2018 predicted years). Y-axis represents MKT_TOTAL_VAL in dollars. First Column represents results of linear regression model; the 2nd column represents quadratic polynomial regression. Each sub-panel in the facet depicts a histogram for the cluster number. TOP: Cluster 1 dominates the downward trend in MKT_TOTAL_VAL. MIDDLE: Clusters 2 dominate increasing trends. Cluster 8 dominates the rising trend .

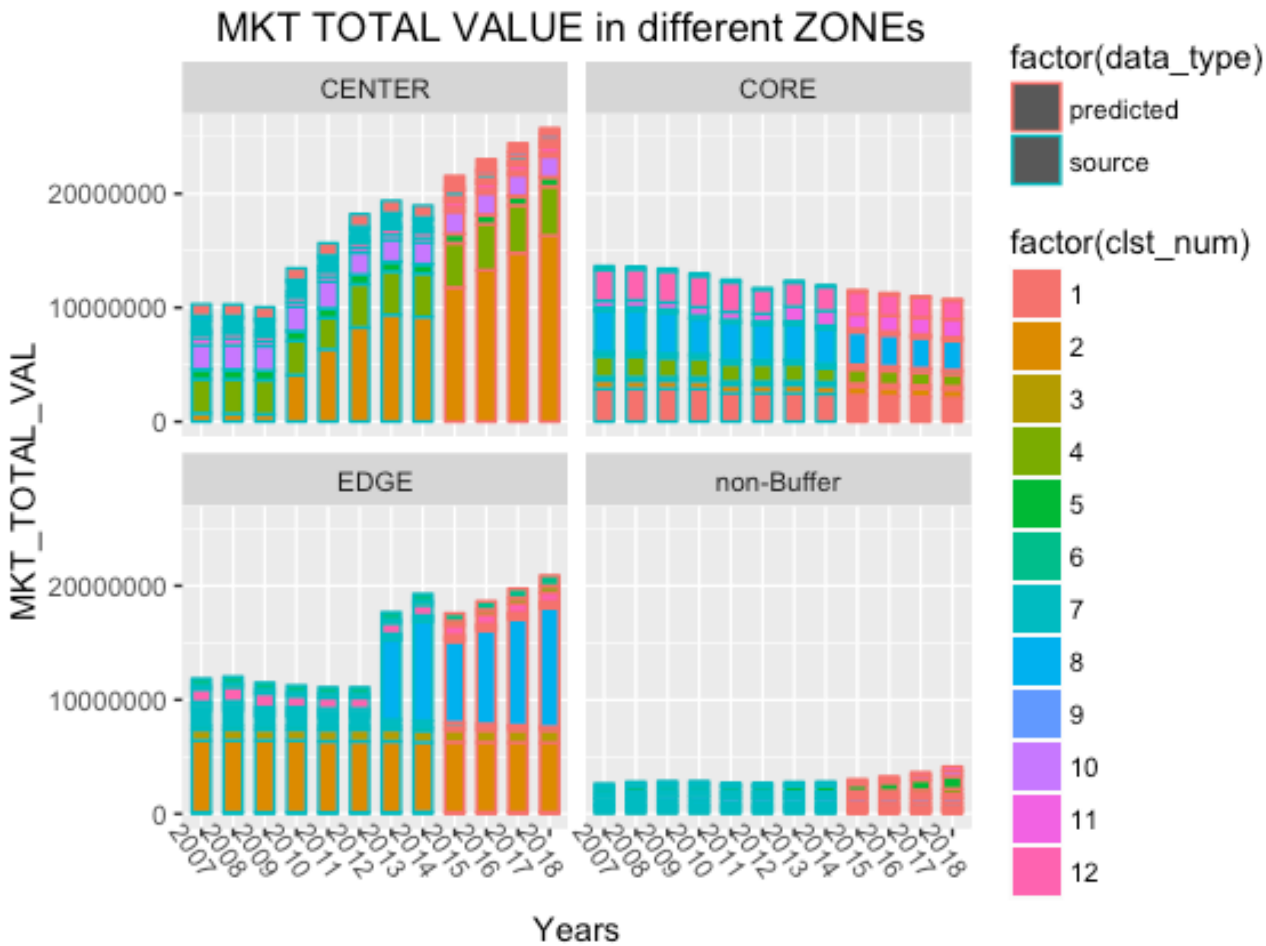


Figure 23: Final comparison of all MKT_TOTAL_VAL. Source data for years 2007-2014 is depicted in histograms with a teal- colored border. The predicted data for years 2015-2018 is depicted in histograms with red-colored border. Of the 4 zones, the CENTER ZONE shows maximum appreciation on value. Surprisingly, CORE zone with closest proximity exhibits declining values. EDGE Zone exhibits increasing market value but is less than the CENTER zone. Finally, the non-buffer zone which is a sample of parcels from the rest of the Hamilton county, shows increasing trends but at much lower rate than the buffer-zone.

Tabulation of Findings

The following Table 9 supports the results of the study as a numerical summary of the findings.

- **zone** - Column 1 is the zone under study *viz*. CORE,CENTER, EDGE, and non-Buffer zones.
- **clst_num** - Within each zone, we had performed clustering and the column 2 , clst_num

is the cluster within the zone.

- **date** - date column depicts the year.
- **data_type -** refers to source data or predicted data.
- **mean_mkt_val** - For each cluster, we had performed row means. Mean_mkt_val represents the mean of all the MKT_TOTAL_VAL for each parcel_id the cluster.
- **num_parcel_ids** - is the number of parcels within the cluster.
- **mkt_total** - is **mean_mkt_val** X **num_parcel_ids**.
- **min_mape_poly** - is the polynomial order of the model with minimum MAPE
- **min_mape** – is the minimum MAPE
-

Table 9: Tabulation of the Source Data 2007-2014 and Predicted data for years 2015-2018

zone	clst_num	date	data_type	mean_mkt_val	num_parcel_ids	mkt_total	min_mape_poly	min_mape
CENTER	1	2007	source	1031284.534	247	254727280	3	0.01634303
		2008	source	1020694.413		252111520		
		2009	source	1020597.935		252087690		
		2010	source	1020597.935		252087690		
		2011	source	1082296.964		267327350		
		2012	source	1082149.352		267290890		
		2013	source	1082128.664		267285780		
		2014	source	1069018.623		264047600		
		2015	predicted	1010939.02		249701937.9		
		2016	predicted	909783.691		224716571.7		
		2017	predicted	753500.9977		186114746.4		
		2018	predicted	532106.2484		131430243.4		
CENTER	2	2007	source	735434.6154	26	19121300	1	0.04892778
		2008	source	705013.0769		18330340		
		2009	source	641629.2308		16682360		
		2010	source	4087940.769		106286460		
		2011	source	6351003.462		165126090		
		2012	source	8262269.231		214819000		
		2013	source	9376237.308		243782170		
		2014	source	9198077.692		239150020		
		2015	predicted	11761822.39		305807382.1		
		2016	predicted	13282293.88		345339641		

zone	clst_num	date	data_type	mean_mkt_val	num_parcel_ids	mkt_total	min_mape_poly	min_mape
		2017	predicted	14802765.38		384871899.8		
		2018	predicted	16323236.87		424404158.6		
CENTER	3	2007	source	149850.4348	46	6893120	2	0.01077346
		2008	source	198189.5652		9116720		
		2009	source	209098.6957		9618540		
		2010	source	177585.6522		8168940		
		2011	source	161628.0435		7434890		
		2012	source	161628.0435		7434890		
		2013	source	161607.6087		7433950		
		2014	source	160826.9565		7398040		
		2015	predicted	133946.875		6161556.25		
		2016	predicted	114506.4868		5267298.393		
		2017	predicted	91807.64881		4223151.845		
		2018	predicted	65850.36102		3029116.607		
CENTER	4	2007	source	2987850	14	41829900	1	0.00843079
		2008	source	2996524.286		41951340		
		2009	source	2996524.286		41951340		
		2010	source	2996524.286		41951340		
		2011	source	2738612.143		38340570		
		2012	source	3789140		53047960		
		2013	source	3789140		53047960		
		2014	source	3788982.143		53045750		
		2015	predicted	3886712.423		54413973.93		
		2016	predicted	4025890.264		56362463.69		
		2017	predicted	4165068.104		58310953.45		
		2018	predicted	4304245.944		60259443.21		
CENTER	5	2007	source	836345.4348	46	38471890	1	0.00123556
		2008	source	839603.6957		38621770		
		2009	source	839603.6957		38621770		
		2010	source	839603.6957		38621770		
		2011	source	836827.6087		38494070		
		2012	source	836827.6087		38494070		
		2013	source	836827.6087		38494070		
		2014	source	804486.087		37006360		

zone	clst_num	date	data_type	mean_mkt_val	num_parcel_ids	mkt_total	min_mape_poly	min_mape
		2015	predicted	820479.9534		37742077.86		
		2016	predicted	817527.5699		37606268.21		
		2017	predicted	814575.1863		37470458.57		
		2018	predicted	811622.8028		37334648.93		
		2007	source	277924.1935		8615650		
		2008	source	278802.5806		8642880		
		2009	source	278802.5806		8642880		
		2010	source	278802.5806		8642880		
		2011	source	280430.9677		8693360		
		2012	source	277895.4839		8614760		
CENTER	6	2013	source	277895.4839	31	8614760	2	0.10045817
		2014	source	281640.3226		8730850		
		2015	predicted	280775.9217		8704053.571		
		2016	predicted	281458.341		8725208.571		
		2017	predicted	282228.7097		8749090		
		2018	predicted	283087.0276		8775697.857		
		2007	source	805675.625		38672430		
		2008	source	806418.125		38708070		
		2009	source	806418.125		38708070		
		2010	source	808501.4583		38808070		
		2011	source	742311.0417		35630930		
		2012	source	742367.0833		35633620		
CENTER	7	2013	source	742367.0833	48	35633620	2	0.05407778
		2014	source	615970.7292		29566595		
		2015	predicted	576251.1365		27660054.55		
		2016	predicted	499976.4317		23998868.72		
		2017	predicted	412986.1502		19823335.21		
		2018	predicted	315280.292		15133454.02		
		2007	source	367359.0476		7714540		
		2008	source	223589.0476		4695370		
		2009	source	223589.0476		4695370		
CENTER	8	2010	source	223589.0476	21	4695370	2	0.11123225
		2011	source	222735.7143		4677450		
		2012	source	222735.7143		4677450		

zone	clst_num	date	data_type	mean_mkt_val	num_parcel_ids	mkt_total	min_mape_poly	min_mape
		2013	source	222735.7143		4677450		
		2014	source	219206.6667		4603340		
		2015	predicted	272374.5238		5719865		
		2016	predicted	318370.7937		6685786.667		
		2017	predicted	376053.8095		7897130		
		2018	predicted	445423.5714		9353895		
CENTER	9	2007	source	35131.25	16	562100	1	0.01659657
		2008	source	34783.75		556540		
		2009	source	34783.75		556540		
		2010	source	34783.75		556540		
		2011	source	34981.875		559710		
		2012	source	34981.875		559710		
		2013	source	34981.875		559710		
		2014	source	34767.5		556280		
		2015	predicted	34858.57143		557737.1429		
		2016	predicted	34849.48661		557591.7857		
		2017	predicted	34840.40179		557446.4286		
		2018	predicted	34831.31696		557301.0714		
CENTER	10	2007	source	2124588.889	9	19121300	1	0.32911508
		2008	source	2124803.333		19123230		
		2009	source	2124347.778		19119130		
		2010	source	2124347.778		19119130		
		2011	source	2326575.556		20939180		
		2012	source	1912751.111		17214760		
		2013	source	1913195.556		17218760		
		2014	source	1904954.444		17144590		
		2015	predicted	1907229.008		17165061.07		
		2016	predicted	1871180.886		16840627.98		
		2017	predicted	1835132.765		16516194.88		
		2018	predicted	1799084.643		16191761.79		
CENTER	11	2007	source	607994.0909	22	13375870	1	0.10145959
		2008	source	606361.8182		13339960		
		2009	source	410733.1818		9036130		
		2010	source	410733.1818		9036130		

zone	clst_num	date	data_type	mean_mkt_val	num_parcel_ids	mkt_total	min_mape_poly	min_mape
		2011	source	439728.1818		9674020		
		2012	source	439728.1818		9674020		
		2013	source	439728.1818		9674020		
		2014	source	383577.7273		8438710		
		2015	predicted	344746.1364		7584415		
		2016	predicted	317506.8182		6985150		
		2017	predicted	290267.5		6385885		
		2018	predicted	263028.1818		5786620		
CENTER	12	2007	source	350216.6667	12	4202600	1	0.20988718
		2008	source	451410.8333		5416930		
		2009	source	451410.8333		5416930		
		2010	source	451410.8333		5416930		
		2011	source	442107.5		5305290		
		2012	source	442107.5		5305290		
		2013	source	505929.1667		6071150		
		2014	source	503576.6667		6042920		
		2015	predicted	519890.8036		6238689.643		
		2016	predicted	535472.9266		6425675.119		
		2017	predicted	551055.0496		6612660.595		
		2018	predicted	566637.1726		6799646.071		
CORE	1	2007	source	2891134.472	322	930945300	1	0.00487282
		2008	source	2893453.043		931691880		
		2009	source	2893493.012		931704750		
		2010	source	2874737.05		925665330		
		2011	source	2507056.056		807272050		
		2012	source	2494528.261		803238100		
		2013	source	2497862.981		804311880		
		2014	source	2433730.528		783661230		
		2015	predicted	2324444.791		748471222.9		
		2016	predicted	2244154.873		722617869		
		2017	predicted	2163864.954		696764515.2		
		2018	predicted	2083575.035		670911161.4		
CORE	2	2007	source	738644.4444	63	46534600	1	0.00529091
		2008	source	719431.746		45324200		

zone	clst_num	date	data_type	mean_mkt_val	num_parcel_ids	mkt_total	min_mape_poly	min_mape
		2009	source	718410.3175		45259850		
		2010	source	703996.3492		44351770		
		2011	source	710781.2698		44779220		
		2012	source	716653.0159		45149140		
		2013	source	729750.6349		45974290		
		2014	source	729263.6508		45943610		
		2015	predicted	720193.6735		45372201.43		
		2016	predicted	720044.1723		45362782.86		
		2017	predicted	719894.6712		45353364.29		
		2018	predicted	719745.1701		45343945.71		
CORE	3	2007	source	289198.0342	117	33836170	1	0.02946052
		2008	source	291259.8291		34077400		
		2009	source	291012.5641		34048470		
		2010	source	291012.5641		34048470		
		2011	source	287659.6581		33656180		
		2012	source	287370.0855		33622300		
		2013	source	287455.3846		33632280		
		2014	source	282704.4103		33076416		
		2015	predicted	284239.8913		33256067.29		
		2016	predicted	283302.2969		33146368.74		
		2017	predicted	282364.7025		33036670.19		
		2018	predicted	281427.1081		32926971.64		
CORE	4	2007	source	1821732.759	58	105660500	1	0.18156008
		2008	source	1822946.724		105730910		
		2009	source	1674670.862		97130910		
		2010	source	1666050.172		96630910		
		2011	source	1528502.931		88653170		
		2012	source	1528502.931		88653170		
		2013	source	1528502.931		88653170		
		2014	source	1231506.207		71427360		
		2015	predicted	1269238.233		73615817.5		
		2016	predicted	1195668.52		69348774.17		
		2017	predicted	1122098.807		65081730.83		
		2018	predicted	1048529.095		60814687.5		

zone	clst_num	date	data_type	mean_mkt_val	num_parcel_ids	mkt_total	min_mape_poly	min_mape
CORE	5	2007	source	318449.3939	33	10508830	2	0.0288083
		2008	source	237923.3333		7851470		
		2009	source	237923.3333		7851470		
		2010	source	237923.3333		7851470		
		2011	source	216569.697		7146800		
		2012	source	145327.5758		4795810		
		2013	source	146449.3939		4832830		
		2014	source	145841.2121		4812760		
		2015	predicted	124490.9416		4108201.071		
		2016	predicted	113731.1508		3753127.976		
		2017	predicted	105497.4206		3481414.881		
		2018	predicted	99789.75108		3293061.786		
CORE	6	2007	source	316585.875	80	25326870	1	0.04827739
		2008	source	317390.875		25391270		
		2009	source	317390.875		25391270		
		2010	source	317390.875		25391270		
		2011	source	318178.25		25454260		
		2012	source	320823.125		25665850		
		2013	source	321368.875		25709510		
		2014	source	319980		25598400		
		2015	predicted	321570.7188		25725657.5		
		2016	predicted	322222.3021		25777784.17		
		2017	predicted	322873.8854		25829910.83		
		2018	predicted	323525.4688		25882037.5		
CORE	7	2007	source	63692.85714	28	1783400	1	0.0059707
		2008	source	66258.92857		1855250		
		2009	source	66258.92857		1855250		
		2010	source	66319.28571		1856940		
		2011	source	55953.57143		1566700		
		2012	source	53012.85714		1484360		
		2013	source	53012.85714		1484360		
		2014	source	70772.14286		1981620		
		2015	predicted	58332.71684		1633316.071		
		2016	predicted	57537.72534		1611056.31		

zone	clst_num	date	data_type	mean_mkt_val	num_parcel_ids	mkt_total	min_mape_poly	min_mape
		2017	predicted	56742.73384		1588796.548		
		2018	predicted	55947.74235		1566536.786		
CORE	8	2007	source	3594574.286	35	125810100	1	0.12056356
		2008	source	3629570.571		127034970		
		2009	source	3574790.857		125117680		
		2010	source	3109410.857		108829380		
		2011	source	3244181.143		113546340		
		2012	source	3087052.286		108046830		
		2013	source	3087052.286		108046830		
		2014	source	3053866.286		106885320		
		2015	predicted	2878312.704		100740944.6		
		2016	predicted	2785146.122		97480114.29		
		2017	predicted	2691979.541		94219283.93		
		2018	predicted	2598812.959		90958453.57		
CORE	9	2007	source	89912.5	20	1798250	1	0.02362266
		2008	source	104987		2099740		
		2009	source	104987		2099740		
		2010	source	93240		1864800		
		2011	source	94524		1890480		
		2012	source	94524		1890480		
		2013	source	94524		1890480		
		2014	source	93927.5		1878550		
		2015	predicted	93418.51786		1868370.357		
		2016	predicted	92771.91071		1855438.214		
		2017	predicted	92125.30357		1842506.071		
		2018	predicted	91478.69643		1829573.929		
CORE	10	2007	source	193699.0909	22	4261380	1	0.04730257
		2008	source	198428.1818		4365420		
		2009	source	219726.3636		4833980		
		2010	source	198036.3636		4356800		
		2011	source	197755.9091		4350630		
		2012	source	198775.4545		4373060		
		2013	source	198775.4545		4373060		
		2014	source	194593.6364		4281060		

zone	clst_num	date	data_type	mean_mkt_val	num_parcel_ids	mkt_total	min_mape_poly	min_mape
		2015	predicted	197020.1461		4334443.214		
		2016	predicted	196363.7771		4320003.095		
		2017	predicted	195707.408		4305562.976		
		2018	predicted	195051.039		4291122.857		
CORE	11	2007	source	601300	19	11424700	1	0.42963005
		2008	source	603130		11459470		
		2009	source	603130		11459470		
		2010	source	734551.0526		13956470		
		2011	source	687368.4211		13060000		
		2012	source	685730.5263		13028880		
		2013	source	1305341.053		24801480		
		2014	source	1317271.053		25028150		
		2015	predicted	1284556.598		24406575.36		
		2016	predicted	1388407.45		26379741.55		
		2017	predicted	1492258.302		28352907.74		
		2018	predicted	1596109.154		30326073.93		
CORE	12	2007	source	2729065.625	16	43665050	1	0.04183377
		2008	source	2713648.75		43418380		
		2009	source	2713273.75		43412380		
		2010	source	2713273.75		43412380		
		2011	source	2599702.5		41595240		
		2012	source	2112985		33807760		
		2013	source	2113215.625		33811450		
		2014	source	2102383.75		33638140		
		2015	predicted	1976298.438		31620775		
		2016	predicted	1865543.958		29848703.33		
		2017	predicted	1754789.479		28076631.67		
		2018	predicted	1644035		26304560		
EDGE	1	2007	source	120247.8723	94	11303300	1	0.13026623
		2008	source	120556.8085		11332340		
		2009	source	120556.8085		11332340		
		2010	source	120556.8085		11332340		
		2011	source	116322.8723		10934350		
		2012	source	116342.8723		10936230		

zone	clst_num	date	data_type	mean_mkt_val	num_parcel_ids	mkt_total	min_mape_poly	min_mape
		2013	source	116342.8723		10936230		
		2014	source	154112.4468		14486570		
		2015	predicted	133796.345		12576856.43		
		2016	predicted	136166.6616		12799666.19		
		2017	predicted	138536.9782		13022475.95		
		2018	predicted	140907.2948		13245285.71		
EDGE	2	2007	source	6304562	65	409796530	1	0.03599647
		2008	source	6308844.462		410074890		
		2009	source	6308844.462		410074890		
		2010	source	6308627.846		410060810		
		2011	source	6233607.846		405184510		
		2012	source	6234028		405211820		
		2013	source	6234028		405211820		
		2014	source	6106907.231		396948970		
		2015	predicted	6144727.566		399407291.8		
		2016	predicted	6120237.863		397815461.1		
		2017	predicted	6095748.159		396223630.4		
		2018	predicted	6071258.456		394631799.6		
EDGE	3	2007	source	944475.6089	271	255952890	1	0.01665143
		2008	source	945766.4945		256302720		
		2009	source	934579.9631		253271170		
		2010	source	947900		256880900		
		2011	source	1007061.365		272913630		
		2012	source	1026780.148		278257420		
		2013	source	1025072.731		277794710		
		2014	source	1012839.815		274479590		
		2015	predicted	1045426.081		283310467.9		
		2016	predicted	1059840.873		287216876.5		
		2017	predicted	1074255.665		291123285.2		
		2018	predicted	1088670.457		295029693.9		
EDGE	4	2007	source	315583.3333	42	13254500	3	0.07392193
		2008	source	316076.6667		13275220		
		2009	source	313424.2857		13163820		
		2010	source	306520.9524		12873880		

zone	clst_num	date	data_type	mean_mkt_val	num_parcel_ids	mkt_total	min_mape_poly	min_mape
		2011	source	317347.8571		13328610		
		2012	source	300665.4762		12627950		
		2013	source	300665.4762		12627950		
		2014	source	328994.2857		13817760		
		2015	predicted	355236.9218		14919950.71		
		2016	predicted	403021.7063		16926911.67		
		2017	predicted	471776.9331		19814631.19		
		2018	predicted	564953.5652		23728049.74		
EDGE	5	2007	source	321278.0645	31	9959620	1	0.00364552
		2008	source	321434.1935		9964460		
		2009	source	321540.3226		9967750		
		2010	source	321540.3226		9967750		
		2011	source	308633.2258		9567630		
		2012	source	308633.2258		9567630		
		2013	source	308633.2258		9567630		
		2014	source	317955.1613		9856610		
		2015	predicted	308765.2419		9571722.5		
		2016	predicted	307111.7473		9520464.167		
		2017	predicted	305458.2527		9469205.833		
		2018	predicted	303804.7581		9417947.5		
EDGE	6	2007	source	632184.2857	119	75229930	3	0.10220159
		2008	source	633349.5798		75368600		
		2009	source	633361.9328		75370070		
		2010	source	636595.4622		75754860		
		2011	source	605891.1765		72101050		
		2012	source	619348.8235		73702510		
		2013	source	665364.4538		79178370		
		2014	source	655418.0672		77994750		
		2015	predicted	703155.1861		83675467.14		
		2016	predicted	758578.1212		90270796.43		
		2017	predicted	834128.8195		99261329.52		
		2018	predicted	932727.9888		110994630.7		
EDGE	7	2007	source	1363196.429	28	38169500	1	0.14268524
		2008	source	1366376.429		38258540		

zone	clst_num	date	data_type	mean_mkt_val	num_parcel_ids	mkt_total	min_mape_poly	min_mape
		2009	source	843081.0714		23606270		
		2010	source	841872.1429		23572420		
		2011	source	794465.7143		22245040		
		2012	source	794465.7143		22245040		
		2013	source	556438.9286		15580290		
		2014	source	555331.0714		15549270		
		2015	predicted	359153.5714		10056300		
		2016	predicted	241320.2679		6756967.5		
		2017	predicted	123486.9643		3457635		
		2018	predicted	5653.660714		158302.5		
EDGE	8	2007	source	201705.6522	23	4639230	1	0.00808937
		2008	source	204030.8696		4692710		
		2009	source	206443.913		4748210		
		2010	source	201504.7826		4634610		
		2011	source	213109.1304		4901510		
		2012	source	195258.6957		4490950		
		2013	source	6970018.261		160310420		
		2014	source	8676753.478		199565330		
		2015	predicted	7097888.121		163251426.8		
		2016	predicted	8206618.126		188752216.9		
		2017	predicted	9315348.131		214253007		
		2018	predicted	10424078.14		239753797.1		
EDGE	9	2007	source	184545.5	20	3690910	1	0.9767113
		2008	source	199257.5		3985150		
		2009	source	199257.5		3985150		
		2010	source	191055.5		3821110		
		2011	source	190544.5		3810890		
		2012	source	190544.5		3810890		
		2013	source	190544.5		3810890		
		2014	source	187643		3752860		
		2015	predicted	189074.1071		3781482.143		
		2016	predicted	188496.3393		3769926.786		
		2017	predicted	187918.5714		3758371.429		
		2018	predicted	187340.8036		3746816.071		

zone	clst_num	date	data_type	mean_mkt_val	num_parcel_ids	mkt_total	min_mape_poly	min_mape
EDGE	10	2007	source	179212.0755	53	9498240	1	0.03259193
		2008	source	179409.0566		9508680		
		2009	source	184932.8302		9801440		
		2010	source	187111.3208		9916900		
		2011	source	182180.1887		9655550		
		2012	source	182180.1887		9655550		
		2013	source	182651.5094		9680530		
		2014	source	179405.283		9508480		
		2015	predicted	182369.717		9665595		
		2016	predicted	182421.8082		9668355.833		
		2017	predicted	182473.8994		9671116.667		
		2018	predicted	182525.9906		9673877.5		
EDGE	11	2007	source	191423	30	5742690	1	0.08242064
		2008	source	222090		6662700		
		2009	source	222090		6662700		
		2010	source	222090		6662700		
		2011	source	195361.6667		5860850		
		2012	source	195361.6667		5860850		
		2013	source	195361.6667		5860850		
		2014	source	196605.3333		5898160		
		2015	predicted	194104.4167		5823132.5		
		2016	predicted	191672.5278		5750175.833		
		2017	predicted	189240.6389		5677219.167		
		2018	predicted	186808.75		5604262.5		
EDGE	12	2007	source	1108875.6	25	27721890	2	0.05512297
		2008	source	1214448		30361200		
		2009	source	1214448		30361200		
		2010	source	974144.8		24353620		
		2011	source	933741.2		23343530		
		2012	source	933741.2		23343530		
		2013	source	933741.2		23343530		
		2014	source	917436.4		22935910		
		2015	predicted	865669.9214		21641748.04		
		2016	predicted	843238.3214		21080958.04		

zone	clst_num	date	data_type	mean_mkt_val	num_parcel_ids	mkt_total	min_mape_poly	min_mape
		2017	predicted	824954.05		20623851.25		
		2018	predicted	810817.1071		20270427.68		
non-Buffer	1	2007	source	185895.6992	379	70454470	1	0.03612288
		2008	source	183416.4644		69514840		
		2009	source	181349.0501		68731290		
		2010	source	181435.0132		68763870		
		2011	source	160547.9578		60847676		
		2012	source	160609.5145		60871006		
		2013	source	160597.1662		60866326		
		2014	source	157388.8654		59650380		
		2015	predicted	150150.5026		56907040.5		
		2016	predicted	145427.2885		55116942.33		
		2017	predicted	140704.0743		53326844.17		
		2018	predicted	135980.8602		51536746		
non-Buffer	2	2007	source	268186.7391	138	37009770	1	0.01456209
		2008	source	254290.5072		35092090		
		2009	source	254311.5217		35094990		
		2010	source	254383.6957		35104950		
		2011	source	238649.5652		32933640		
		2012	source	238649.5652		32933640		
		2013	source	234742.1014		32394410		
		2014	source	222557.6087		30712950		
		2015	predicted	220014.309		30361974.64		
		2016	predicted	214301.6192		29573623.45		
		2017	predicted	208588.9294		28785272.26		
		2018	predicted	202876.2396		27996921.07		
non-Buffer	3	2007	source	182612.3404	94	17165560	1	0.01231621
		2008	source	207747.3404		19528250		
		2009	source	205558.7234		19322520		
		2010	source	205665.4255		19332550		
		2011	source	178489.2553		16777990		
		2012	source	176902.9787		16628880		
		2013	source	176902.9787		16628880		
		2014	source	177940.4255		16726400		

zone	clst_num	date	data_type	mean_mkt_val	num_parcel_ids	mkt_total	min_mape_poly	min_mape
		2015	predicted	172902.329		16252818.93		
		2016	predicted	169330.0836		15917027.86		
		2017	predicted	165757.8381		15581236.79		
		2018	predicted	162185.5927		15245445.71		
non-Buffer	4	2007	source	214978.6709	158	33966630	3	0.02227791
		2008	source	221680.1266		35025460		
		2009	source	220773.1013		34882150		
		2010	source	220620.3797		34858020		
		2011	source	187020		29549160		
		2012	source	187025.9494		29550100		
		2013	source	187844.3671		29679410		
		2014	source	190407.5316		30084390		
		2015	predicted	218786.1347		34568209.29		
		2016	predicted	267965.8258		42338600.48		
		2017	predicted	344825.434		54482418.57		
		2018	predicted	454428.7779		71799746.9		
non-Buffer	5	2007	source	365227.027	37	13513400	3	0.12223967
		2008	source	374416.7568		13853420		
		2009	source	374416.7568		13853420		
		2010	source	374416.7568		13853420		
		2011	source	371042.973		13728590		
		2012	source	378574.5946		14007260		
		2013	source	457101.3514		16912750		
		2014	source	484612.1622		17930650		
		2015	predicted	586195.3475		21689227.86		
		2016	predicted	715658.0824		26479349.05		
		2017	predicted	888182.0785		32862736.9		
		2018	predicted	1109877.782		41065477.94		
non-Buffer	6	2007	source	166446.4444	90	14980180	1	0.01472761
		2008	source	169183.1111		15226480		
		2009	source	169183.1111		15226480		
		2010	source	169183.1111		15226480		
		2011	source	165530.7778		14897770		
		2012	source	165530.7778		14897770		

zone	clst_num	date	data_type	mean_mkt_val	num_parcel_ids	mkt_total	min_mape_poly	min_mape
		2013	source	165530.7778		14897770		
		2014	source	161178.6667		14506080		
		2015	predicted	162734.4841		14646103.57		
		2016	predicted	161904.1812		14571376.31		
		2017	predicted	161073.8783		14496649.05		
		2018	predicted	160243.5754		14421921.79		
non-Buffer	7	2007	source	137050	18	2466900	2	0.01727332
		2008	source	144603.3333		2602860		
		2009	source	144603.3333		2602860		
		2010	source	144603.3333		2602860		
		2011	source	145852.7778		2625350		
		2012	source	145852.7778		2625350		
		2013	source	145852.7778		2625350		
		2014	source	137625		2477250		
		2015	predicted	134210.2579		2415784.643		
		2016	predicted	127816.5807		2300698.452		
		2017	predicted	120107.8108		2161940.595		
		2018	predicted	111083.9484		1999511.071		
non-Buffer	8	2007	source	180403.5849	53	9561390	1	0.19031685
		2008	source	284280.3774		15066860		
		2009	source	288860.1887		15309590		
		2010	source	301064.5283		15956420		
		2011	source	279545.4717		14815910		
		2012	source	278516.4151		14761370		
		2013	source	274063.9623		14525390		
		2014	source	271430.3774		14385810		
		2015	predicted	298353.9218		15812757.86		
		2016	predicted	304705.7682		16149405.71		
		2017	predicted	311057.6146		16486053.57		
		2018	predicted	317409.4609		16822701.43		
non-Buffer	9	2007	source	297534.1026	39	11603830	1	0.08927162
		2008	source	297165.8974		11589470		
		2009	source	360857.1795		14073430		
		2010	source	351456.1538		13706790		

zone	clst_num	date	data_type	mean_mkt_val	num_parcel_ids	mkt_total	min_mape_poly	min_mape
		2011	source	342347.4359		13351550		
		2012	source	336486.1538		13122960		
		2013	source	334160.5128		13032260		
		2014	source	340859.2308		13293510		
		2015	predicted	354359.7894		13820031.79		
		2016	predicted	359193.4463		14008544.4		
		2017	predicted	364027.1032		14197057.02		
		2018	predicted	368860.7601		14385569.64		
		2007	source	223050		24089400		
		2008	source	226081.9444		24416850		
		2009	source	225910.5556		24398340		
		2010	source	225991.0185		24407030		
		2011	source	205502.1296		22194230		
		2012	source	205332.037		22175860		
non-Buffer	10	2013	source	205416.0185	108	22184930	3	0.04280428
		2014	source	217306.4815		23469100		
		2015	predicted	243916.8452		26343019.29		
		2016	predicted	289740.8752		31292014.52		
		2017	predicted	358699.2967		38739524.05		
		2018	predicted	454862.4325		49125142.71		
		2007	source	184593.2099		14952050		
		2008	source	189390.9877		15340670		
		2009	source	188686.2963		15283590		
		2010	source	188793.4568		15292270		
		2011	source	180538.3951		14623610		
		2012	source	180834.321		14647580		
non-Buffer	11	2013	source	180995.679	81	14660650	3	0.01586791
		2014	source	195380.7407		15825840		
		2015	predicted	217591.2698		17624892.86		
		2016	predicted	254944.806		20650529.29		
		2017	predicted	309625.8554		25079694.29		
		2018	predicted	384622.5736		31154428.46		
non-Buffer	12	2007	source	251437.931	29	7291700	2	0.00225263
		2008	source	250638.9655		7268530		

zone	clst_num	date	data_type	mean_mkt_val	num_parcel_ids	mkt_total	min_mape_poly	min_mape
		2009	source	250638.9655		7268530		
		2010	source	250638.9655		7268530		
		2011	source	250830		7274070		
		2012	source	250830		7274070		
		2013	source	250830		7274070		
		2014	source	252573.4483		7324630		
		2015	predicted	253159.218		7341617.321		
		2016	predicted	254333.6515		7375675.893		
		2017	predicted	255719.9528		7415878.631		
		2018	predicted	257318.1219		7462225.536		

Part III: Conclusions and Future Work

Conclusions

In the following pages, the table below summarizes the findings of this study categorized by the zones.

One would have thought, the introduction of the streetcar in downtown would have shown a rising trend in the forecast of the MKT_TOTAL_VAL of the CORE buffer zone. Surprisingly, this zone exhibited declining forecasts for years 2015-2018.

<table>
<tr><th>ZONE</th><th>Summary</th></tr>
<tr><td>CORE</td><td>
1. Overall, total Market value for CORE zone showed declining trend
2. 10 out of a total 12 clusters exhibited declining trends
3. 322 parcels of cluster 1 dominated the CORE zone’s declining trends
a. Declining $748.47M in 2015 to $670.9M in 2018 (-10.36% change)
<table>
<tr><td>2015</td><td>$748,471,222.86</td></tr>
<tr><td>2016</td><td>$722,617,869.05</td></tr>
<tr><td>2017</td><td>$696,764,515.24</td></tr>
<tr><td>2018</td><td>$670,911,161.43</td></tr>
</table>
b. Average value / parcel of $2.32M in 2015 declined to $2.06M in 2018
4. 19 Parcels making up cluster 11 showed increasing trends
a. Increasing Forecasts from $24M in 2015 to $30M in 2018 (+24.2%)
b. Average value / parcel of $1.284M in 2015 increasing to $1.596M in 2018
</td></tr>
</table>

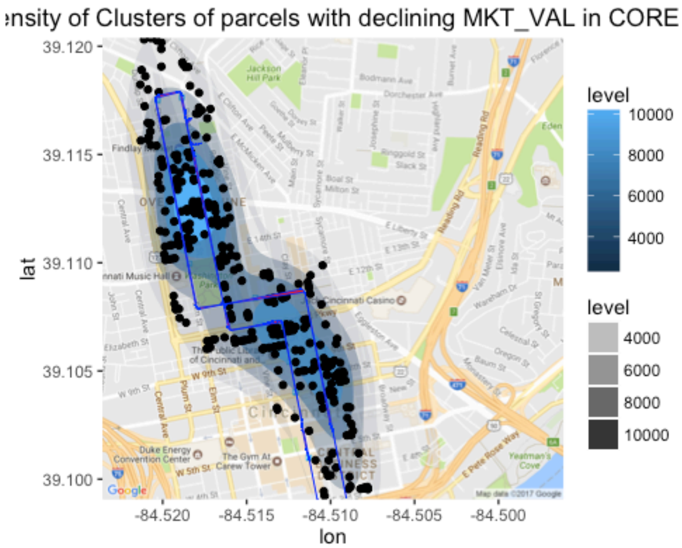


Figure 24: 322 parcels of cluster 1 in the CORE dominate declining trends. There was a 10% decline of MKT_TOTAL_VAL from $748M to $670M. The spatial distribution appears evenly distributed along the route of the streetcar overall. However, there is a dense concentration of parcels bounded by E. Liberty St. and E. 14th St. and between Elm St. and Race St.

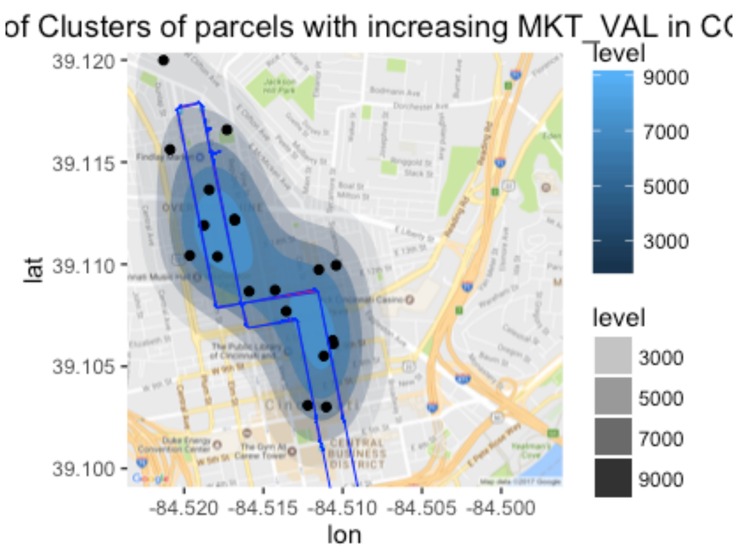


Figure 25: 19 parcels of cluster 11 in CORE show increasing trends. There was a 24.2% increase from $24.4M in 2015 to $30.36M in 2018

CENTER

1. Overall, The Market Total Value for CENTER zone showed upward trend
2. 4 out of a total 12 clusters exhibited increasing trends
3. 26 parcels of cluster 2 dominated the CENTER zone's increasing trends
   a. Increasing from $305.80M in 2015 to $424.404M in 2018 (+38.78% change)

2015	$305,807,382.14
2016	$345,339,640.95
2017	$384,871,899.76
2018	$424,404,158.57

   b. Average value / parcel of $11.761M in 2015 declined to $16.323M in 2018
4. 247 Parcels making up cluster 1 showed declining trends
   a. Forecast predicts a decline in from $249.7M in 2015 to $131.43M in 2018 (-47.37%)
   b. Average value / parcel of $1.01M in 2015 declining to $0.532M in 2018

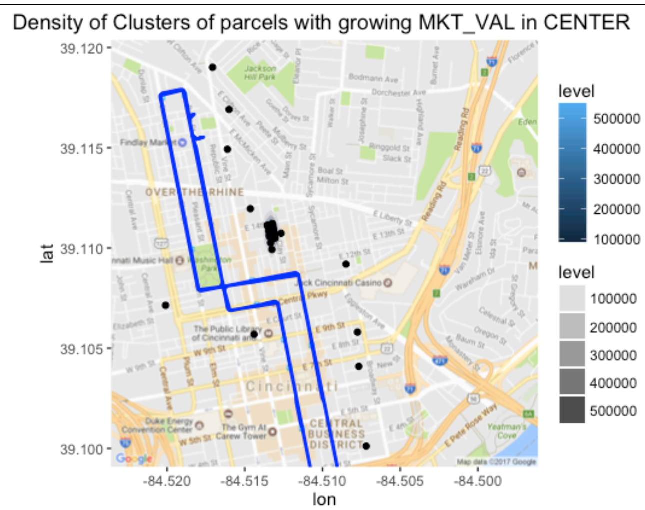


Figure 26: 26 parcels of cluster 2 in the CENTER dominate increasing trends. There was a 38.78% increase of MKT_TOTAL_VAL from $305M to $424M. The spatial distribution appears densely co-located around Walnut St and Clay St. and E 14th and E. 13th St.

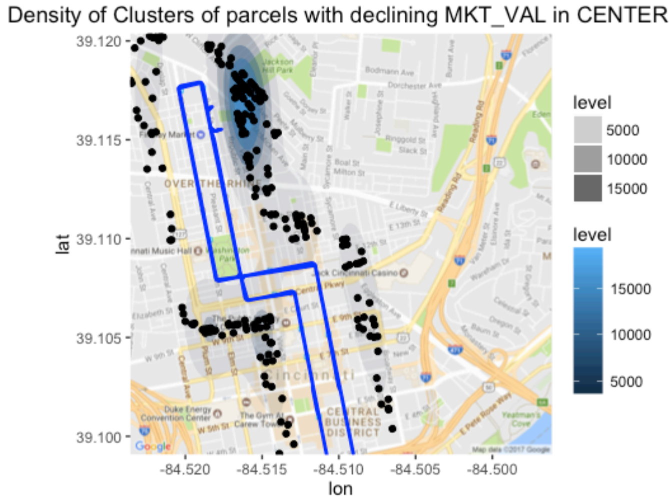


Figure 27: 247 parcels of Cluster 1 show increasing trends. It appears that there maybe a 47% decline from $249M in 2015 to $131M in 2018. There is a dense the 2D_density plot shows a concentration of parcels around Peete St., McMicken Ave. and Clifton Avenue at the North East edge of the Street Car route.

EDGE

5. Overall, total Market value for EDGE zone showed increasing trend
6. 6 out of a total 12 clusters exhibited increasing trends
7. 23 parcels of cluster 8 dominated the EDGE zone's increasing trends
   a. Increasing from $163.251M in 2015 to $239.75 in 2018 (+46% change)

2015	$163,251,426.79
2016	$188,752,216.90
2017	$214,253,007.02
2018	$239,753,797.14

   b. Average value / parcel of $7.097M in 2015 declined to $10.424M in 2018
8. 28 Parcels making up cluster 7 showed declining trends
   a. Forecasts declined from $10.06M in 2015 to $0.158M in 2018 (-98.43%)
   b. Average value / parcel of $359K in 2015 increasing to $5K in 2018

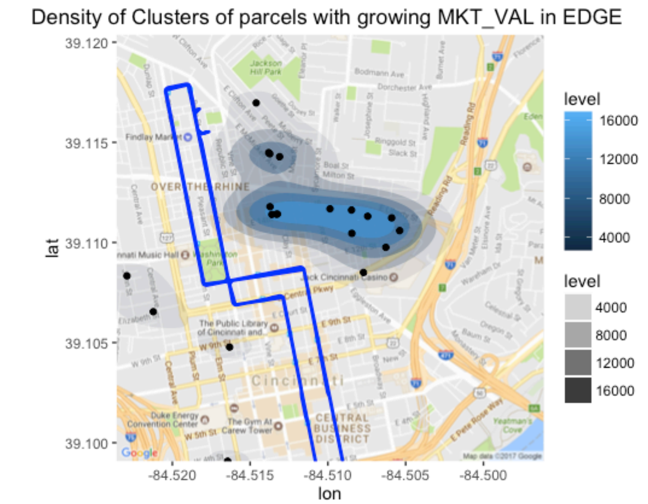


Figure 28: 23 parcels of cluster 8 in the EDGE dominate increasing trends. It seems that there maybe a 46% increase of MKT_TOTAL_VAL from $163M to $239M.

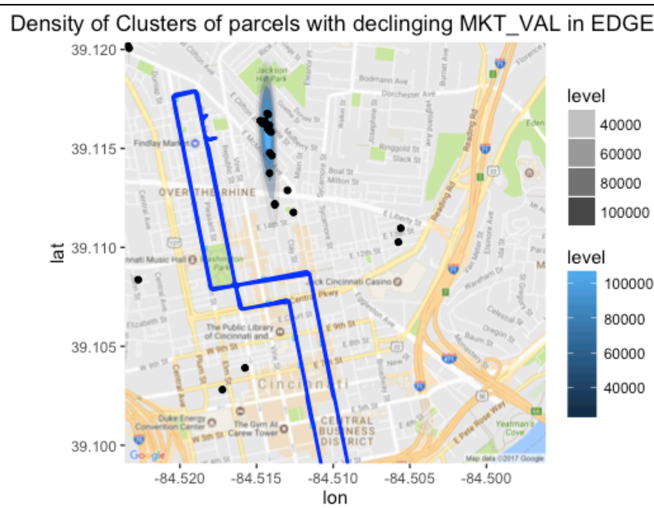


Figure 29: 28 parcels of cluster 7 in EDGE show declining trends. It seems that there may be a 98.43% decline from $1M in 2015 to $0.158M in 2018M. As in the CENTER zone, these appear closer to Clifton Ave.

Summary

Further analyses and data are required to improve the predictive abilities of the regression model. While it appears that the introduction of the streetcar, is accompanied by declining Market Total Value in close proximity of the streetcar route, other factors that may contribute to a City's revenue and the re-vitalization of the area around the Street Car route- ridership, new permits for business, and building improvements has not been factored into the study. The Streetcar was introduced in August 2016. The source-data that has been used for forecast is for years 2007-2014.

Furthermore, the frequency of the data is annual and the fact that the data was collected before the introduction of the streetcar, the prediction may not account for (positive or negative) effects after the event of its introduction.

Future Work

This study centered on the MKT_TOTAL_VAL as the primary variable for study. Further enhancements can be made with the use of other economic indicators, such as MKT_LAND_VAL, MKT_IMPR_VAL, SALE_AMOUNT, NEW_CONSTR flag, ANNUAL_TAXES, TAXES_PAID, Delinquent taxes, Tax foreclosure flag etc. The use of more recent and up-to-date 2015-2016 data would also help develop a more accurate model for prediction.

Acknowledgements

This project is a culmination of my efforts but most importantly the hidden and powerful hand of encouragement and inspiration of many.

I am highly indebted to my mentor - Mr. Anirban Ghosh. M.S. (Machine Learning Specialist, Nokia). This work would not have been possible without his unwavering guidance and supervision. The insights he provided from his knowledge and industry experience were invaluable. His constant encouragement and availability to address any the most simple to most complex problems helped me conclude the project successfully.

I wish to thank Mr. Raj Chundur, Director, Cincinnati Area Geographic Information Systems

(CAGIS) for providing access to the data set and in providing subject matter expertise (Urban planning) and selection of the project objectives.

Last but not the least, I wish thank my family - Ananth, Nithya and wife Dipa for supporting me in taking the first step in laying a new foundation in a career in Data Science.

Bibliography

References

1. https://dev.socrata.com/foundry/data.cincinnati-oh.gov/emnx-rw6d
2. http://www.cincinnati.com/story/news/2016/05/05/streetcar-nation-kc-opens-friday-cincy-next/83874740/
3. https://cran.r-project.org/web/packages/ClustOfVar/ClustOfVar.pdf
4. http://www.exegetic.biz/blog/2013/12/contour-and-density-layers-with-ggmap/
5. http://www.shanelynn.ie/massive-geocoding-with-r-and-google-maps/
6. http://stat405.had.co.nz/ggmap.pdf
7. A Little Book of R for Time-series*, Avril Coghlan*
8. An Introduction to Statistical Learning, *Gareth James et al*

Appendix –A

Problems Encountered during Development of Methodology

Problem: The 9 observations were the annual taxes assessed for years 9 years from 2007-2015. The following problems were then identified:

- Not all parcels had all 9 years of data
- A number of parcels had 0 Annual taxes which could be due to abatement of taxes e.g. a school would finally pay 0 assessed taxes even though the Market value of the parcel on which it was situated was valued higher

Solution: Therefore, Data required some filtering:

- Only those parcels for which all 9 years of data was available were considered: It was decided to remove 23 parcel ids. i.e. 0.0095% of the total parcels.

Further, data with 0 annual taxes was removed.

Problem: A visual inspection, revealed an issue with the data:

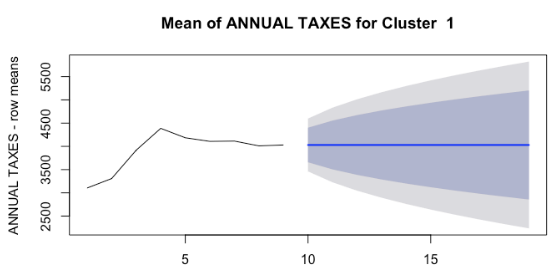


Figure 30: HoltsWinters Simple Exponential smoothing and prediction for cluster 1. The source data mean for years 8 and 9 indicating of problem data set.

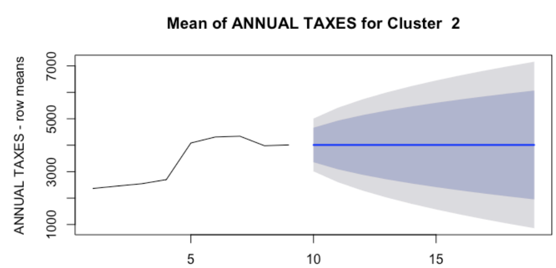


Figure 31: HoltsWinters Simple Exponential smoothing and prediction for cluster 2. The source data mean for years 8 and 9 indicating of problem data set.

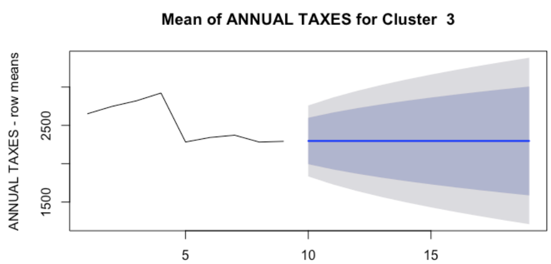


Figure 32: HoltsWinters Simple Exponential smoothing and prediction for cluster 3. The source data mean for years 8 and 9 indicating of problem data set.

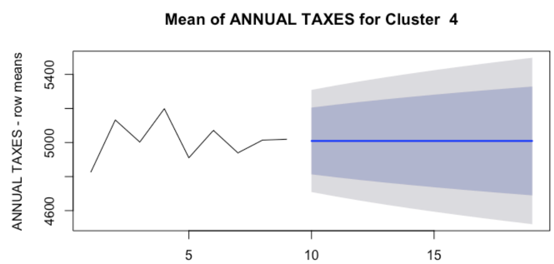


Figure 33: HoltsWinters Simple Exponential smoothing and prediction for cluster 4. The source data mean for years 8 and 9 indicating of problem data set.

For the years 8 and 9, i.e. 2014 and 2015, the taxes were flat. A consultation with the client, CAGIS, also confirmed the observation. It was surmised that batch updates to the Annual Taxes received by the city did not make it to the 2015 tax data .csv files.

Solution: It was decided to use data for 8 years instead of 9 years. Also, it was decided to perform the study on Total Market Value (MKT_TOTAL_VAL was selected as a variable) instead of Annual Taxes.   An annual tax is assessed on the **Total Market Value**. So, selecting this feature would result in **a better indicator** of the "economic effect" as it **removes the effect** of tax **abatements** and **tax incentives** on certain parcels.

Problem: Using HoltWinters on 9 years of annual data

Holt Winters Filtering

If one has a time series that can be described using an additive model with increasing or decreasing trend and seasonality, one can use Holt-Winters exponential smoothing to make short-term forecasts. Holt-Winters exponential smoothing estimates the level, slope and seasonal component at the current time point. Smoothing is controlled by three parameters: alpha, beta, and gamma.

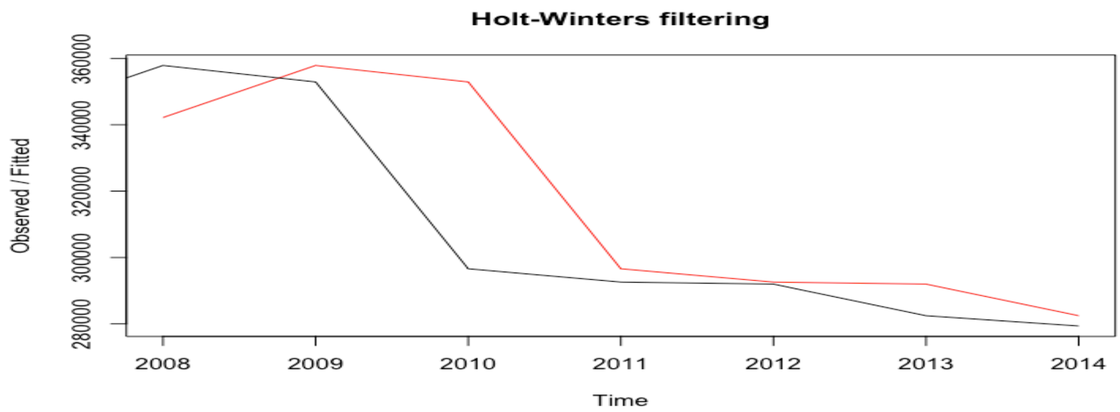


Figure 34: Actual vs. Fitted values of in sample source data. The black line shows the observed in-sample source data for the years 2007 – 2014. The red line marks the fitted data i.e. the prediction of the forecast model.

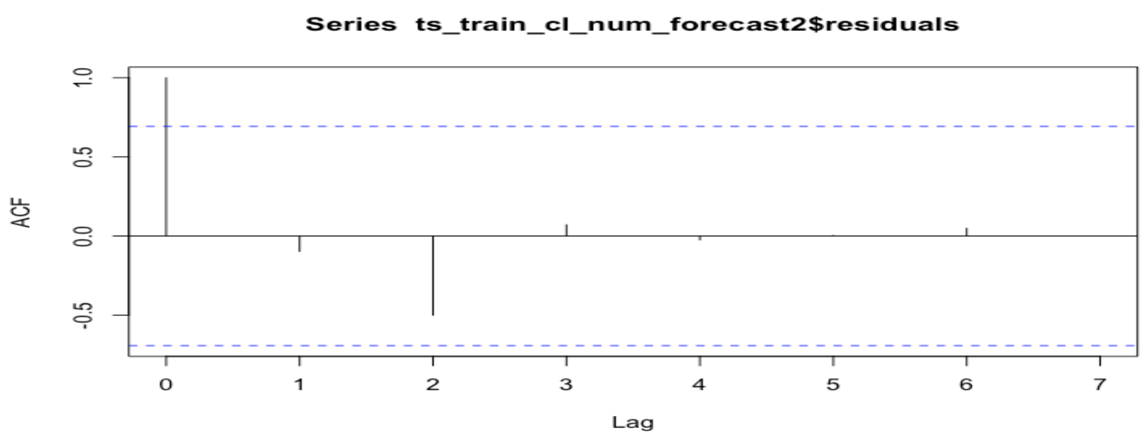
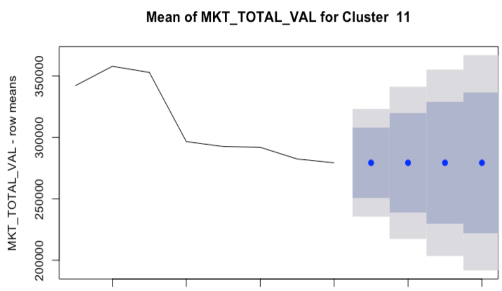


Figure 35: Autocorrelogram indicates that the forecast errors lie within significance bounds. This was the observation for all of the clusters.



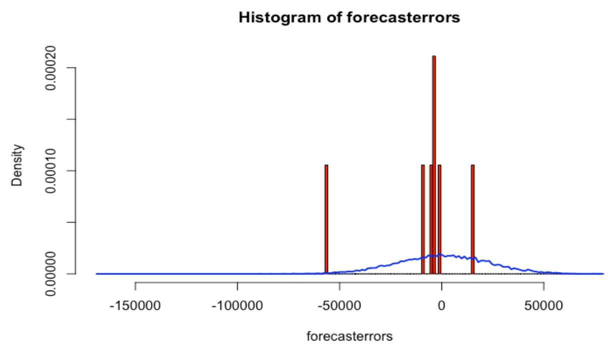


*Figure 36: **Left:** Holts Winters Forecast Source data and Forecast data plotted with the 80% and 95% confidence intervals. **Right**: Histogram of forecast errors shows that the variance is not close to 0*

In figure 35, even though the auto-correlation function ACF lies within the bounds of significance levels, it is observed in figure 36, that the variance of forecast errors is not zero and therefore, the Holts Winters Exponential method was rejected.

Solution: polynomial regression was considered