|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| |  | | --- | |  | |  | |  | |  | |
| Net Economic impact: Introduction of Street Car in Downtown Cincinnati  Exploratory Data Analysis, Predictive Analysis and Forecast |
| **K. Rajesh Jagannath**  **Foundations of Data Science** Mentor: Anirban Ghosh 08/05/2016 |

Table of Contents

Net Economic impact: Introduction of Street Car in Downtown Cincinnati 1

Introduction 1

Objective 1

Motivation 1

Data Sources 2

Extraction, Transformation and Loading of Data 3

Feature Extraction 4

Exploratory Data Analysis 6

Conclusion 7

# Net Economic impact: Introduction of Street Car in Downtown Cincinnati

# Introduction

[The Cincinnati Streetcar](http://www.cincinnati-oh.gov/streetcar/design-route/) 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 positive effect” over 4 more years on the economy in a buffer zone around the streetcar route by selecting Total Market Value of parcels from source data sets.

## Motivation

The City of Cincinnati is the client. Downtown is Cincinnati’s largest employment center, with approximately 70,000 people working in the area every day. 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.** The Annual Taxes assessed on a parcel is a function of Total Market Value of the parcel. Also, here may have been inconveniences to the neighborhood, during the construction phase.  Hence, 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 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

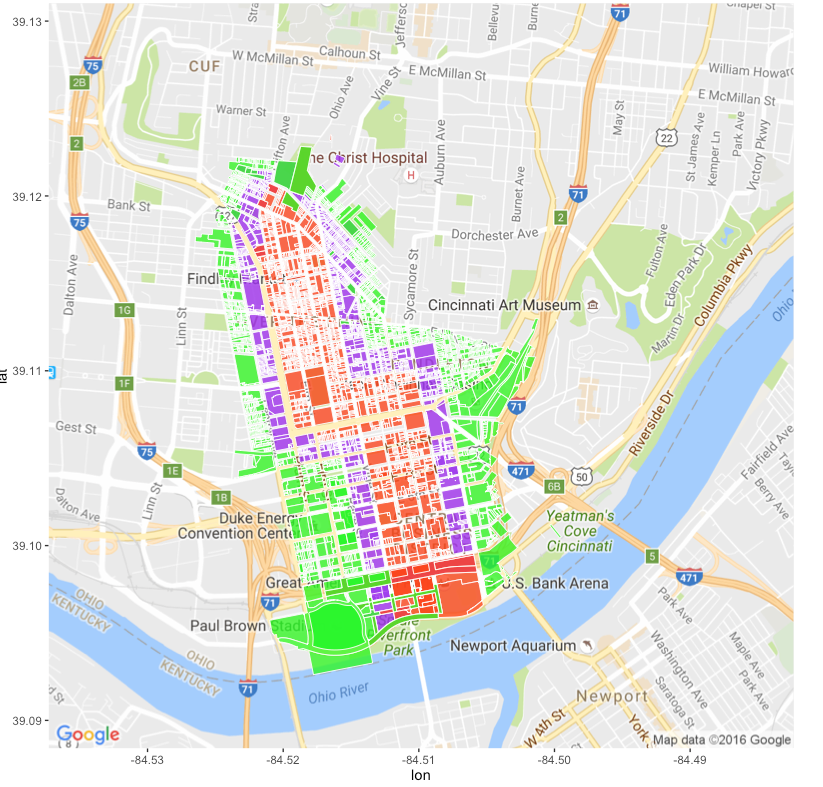


Table . ggplot of the downtown area under study illustrating the Buffer zones CORE, CENTER and EDGE around the Street Car route

# This report is organized in 2 parts. Part 1 concerns itself with Analytics of the Buffer zone. Part 2 covers a Time series forecast of economic impact of the introduction of the streetcar system.

# Part -1: Analysis of the Zones under study

Introduction: Part of this report concerns itself with the analytics of the zone under study. The sources of data, its confidentiality, and problems encountered with data, observations and conclusions are covered in this part

Part - 1: Data and Sources

## **Data 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:** Data is considered to be **Open Data** available to public for use. Permission was obtained further to be able to post the files on github for the purpose of this study

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

* StreetCarParcels\_CORE.csv
* StreetCarParcels\_CENTER.csv
* StreetCarParcels\_EDGE.csv

# Part - 1: Extraction, Transformation and Loading of Data

1. **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 |
| **ADDRST** | SYCAMORE |
| **ADDRSF** | ST |

**Table 2: .csv files are used to identify the parcel id. of the three areas around the Street Car - Core, Center and Edge Buffer zones**

1. 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 = 946 parcels x p = 67 columns
* StreetCarParcels\_CENTER.csv : n = 1418 parcels x p = 67 variables
* StreetCarParcels\_EDGE.csv : n = 1713 parcels x p = 67 variables

# Part1: Exploratory Data Analysis

## Buffer Zones under study: CENTER, CORE and EDGE

Plotting the position - 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 the Figures, 1,2 and 3 below.



Figure . Scatterplot CORE Buffer Zone



Figure . Scatterplot CENTER Buffer Zone



Figure . Scatterplot EDGE Buffer Zone

**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. Also, there are too many parcels classified as vacant lots, which was confirmed by looking at GIS(Geographical information systems) data.

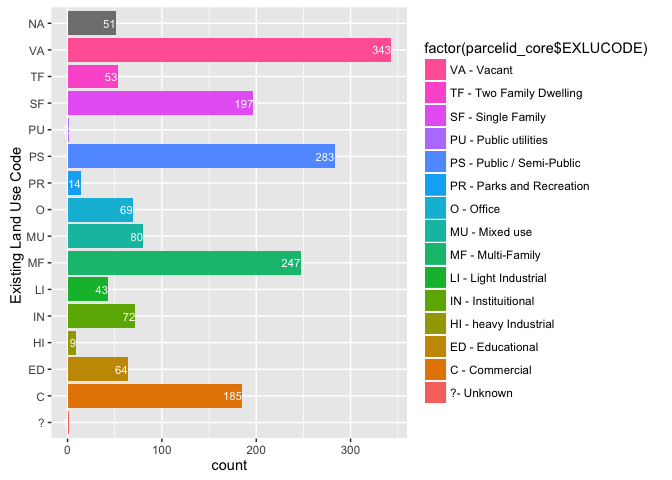


Figure . Histogram of parcels in CORE buffer zone.

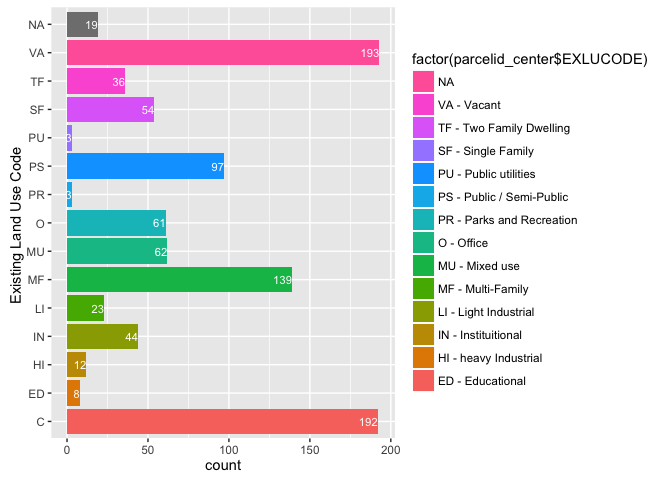


Figure . Histogram of parcels in CENTER buffer zone.

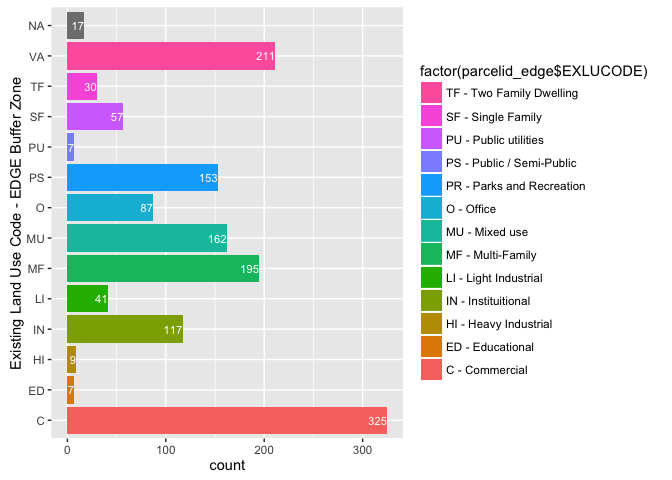


Figure . Histogram of parcels in the EDGE buffer zone

**Methodology:** Performing 2D Kernel Density plot, we find the areas of high Market Land Value to be centered on the Buffer Zone. There **is an unanticipated concentrated distribution in the center of the Downtown** in all the three plots. This is indicative of **a problematic geocoding or the street addresses in the data are not correct**. In the scatterplot, Figure 1, Figure 2 and Figure 3, 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.

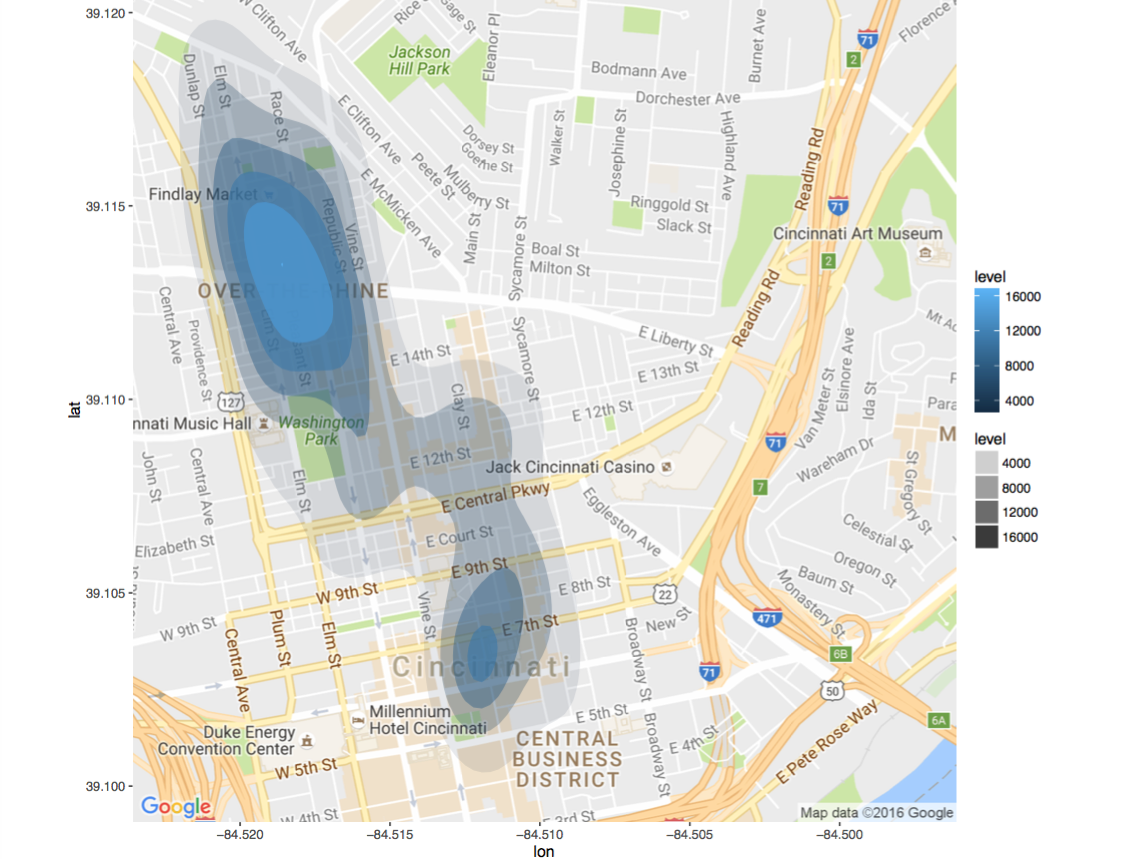


Figure . 2-D Kernel Density plot of CORE

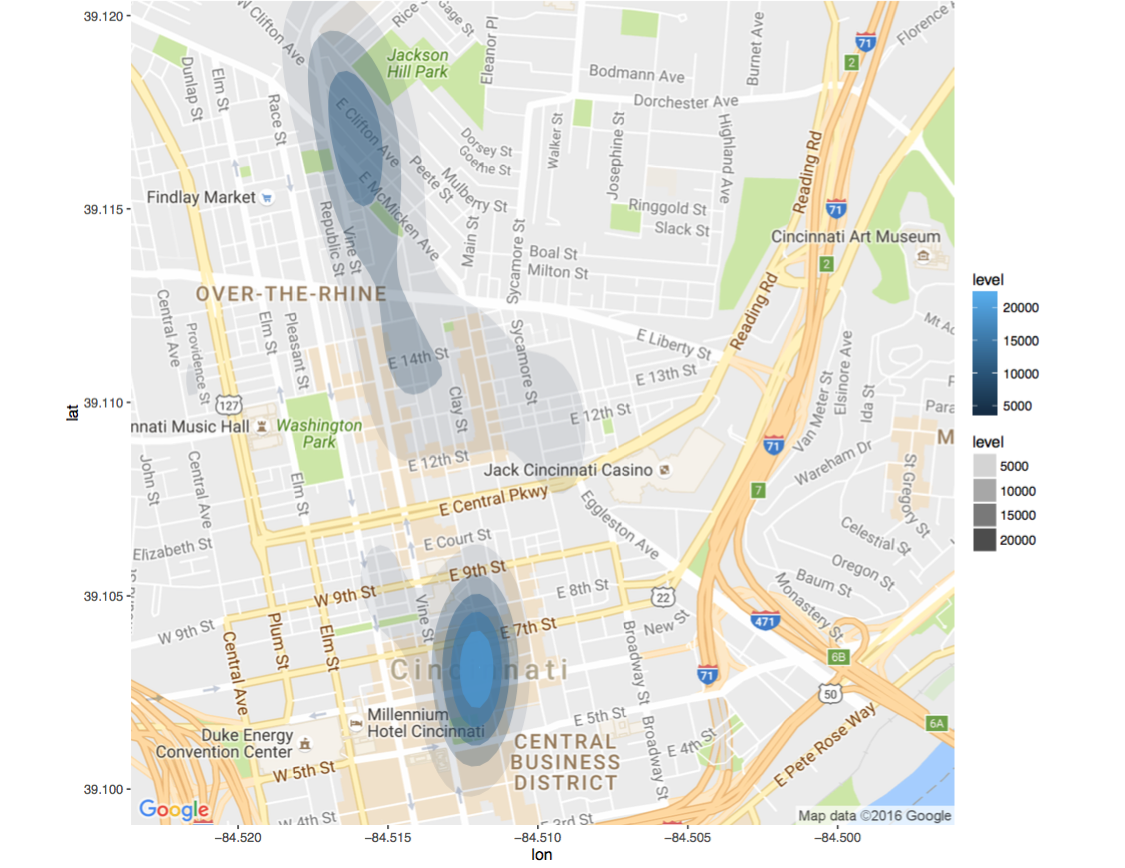


Figure . 2-D Kernel Density plot of CENTER: High density of observations near Central Business District is not expected

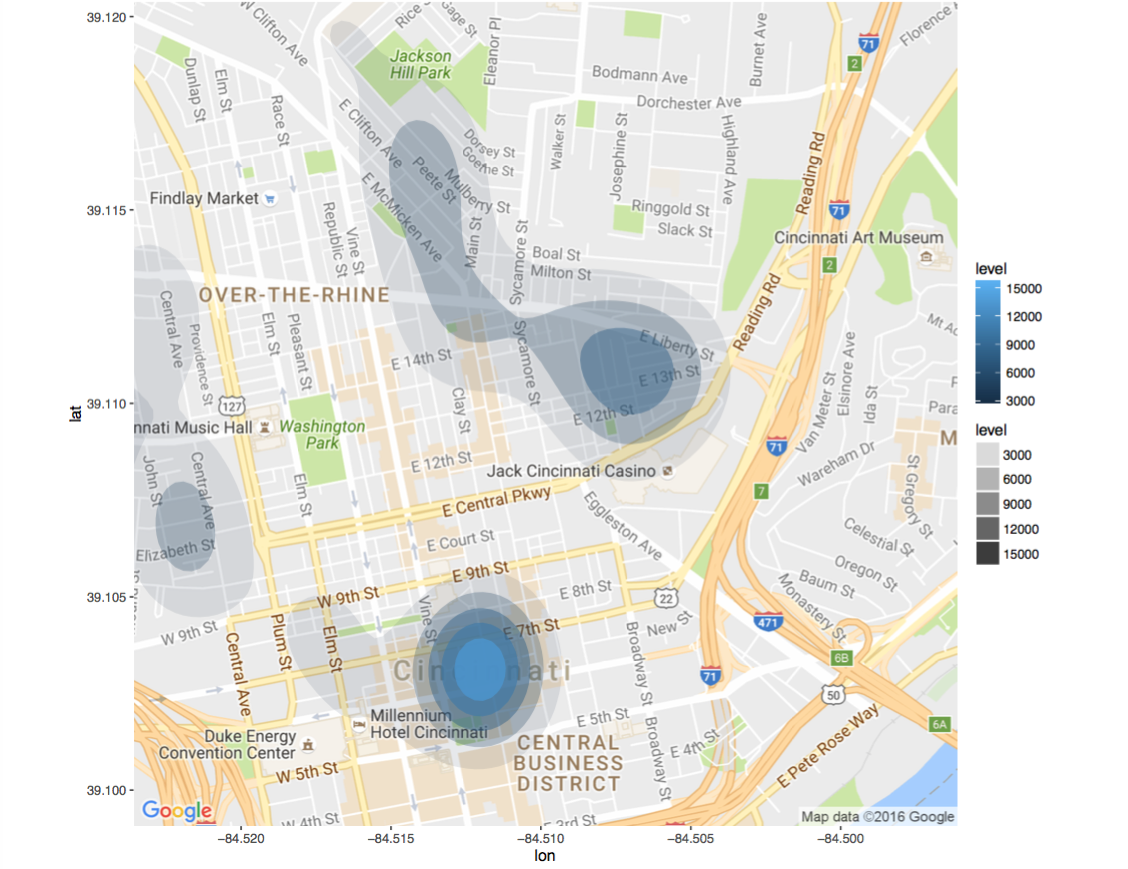


Figure . 2D Kernel Density plots of EDGE: High density of observations near Central Business District is not expected

The data sets were analyzed further with CAGIS and a cleaner data set was obtained

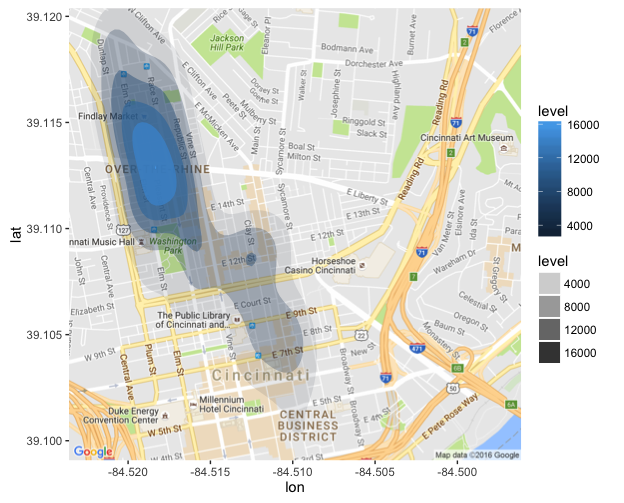


Figure . 2-D Kernel Density plot of the CORE parcels : AFTER – Clean Dataset provided by CAGIS

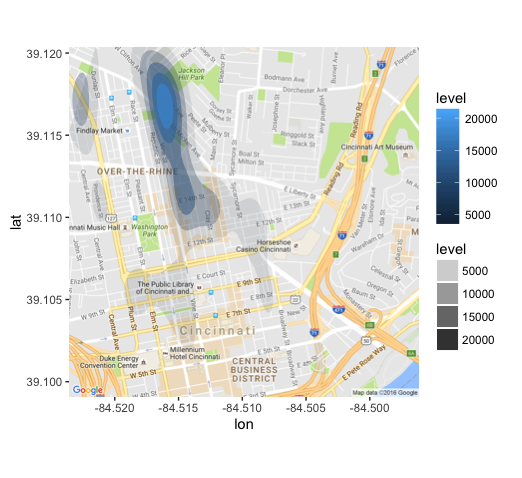


Figure . 2-D Kernel Density plot of the CENTER parcels: AFTER - Clean data provided by CAGIS - Central Business District are no longer there

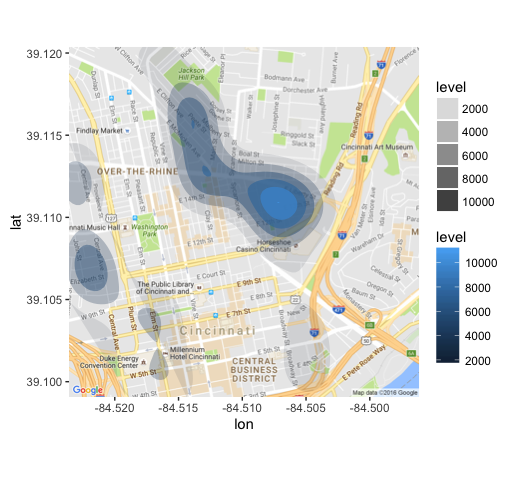


Figure . 2-D Kernel Density plots of the EDGE : After data was cleaned, observations are in line with expectatio

# Part-1: Conclusion

* 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 0 Annual Taxes value were analyzed further. They correspond to one of the several parcels owned by the same owner. The taxes are assessed on one parcel only for billing convenience and others are marked 0.
* Density map indicated some geocoded co-ordinates are not spatially situated in the buffer zones as expected.
  + For example, in Fig. 8 and Fig. 9, there is a high density of observations Near the Central Business District which seems 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
* More accurate data-set was requested
  + Fig. 10, Fig. 11, Fig. 12 illustrates a better distribution of the parcels in the expected buffer zones
  + In particular, the high density of observations near the Central Business District in Fig. 7, 8 and 9, prior to clean up is no longer observed
  + This paves way for sub-setting data for Forecast Analysis ( Part 2)
  + Instead of using Google Maps geocoding, it was decided to obtain longitude and latitude co-ordinates from CAGIS directly

# Part - 2: Economic impact – Time series Forecast

Part - 2: Data and Sources

## **Data Sources**

**Data sources** were Cincinnati Area Geographic Information Systems (CAGIS), City of Cincinnati, OH and in turn from the Hamilton County Auditor Office, Cincinnati, OH.

**Confidentiality of data:** Data is considered to be **Open Data** available to public for use. Data is real and not from any competition or academic study. Permission was obtained further to be able to post the files on github for the purpose of this study.

1. **Assessors Tax Information 2007-2015: The Assessors Office**  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** |  |
| **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 9. Features selected from Property Tax Information from years 2007 - 2015**

# Part-2: Extraction, Transformation and Loading of Data

1. **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 300,000 observations for each year. Files are about in size for each year. The final loaded data frame dimensions were 2,190,994 x 13

# Part-2: Feature Selection

From the tax-information of 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** |
| 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 10. 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 300,000 observations/year | | | | | | | | | |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |

**Table 11. 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.**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **PARCELID** | **YEAR** | **LOCATION** | **EXLU** | **MKT VALUE** | **TAXES** | **SALE** |
|  | 2007 |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  | 2007 |  |  |  |  |  |
|  | | | | | | |
|  | 2008 |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  | 2008 |  |  |  |  |  |
|  | | | | | | |
|  | 2015 |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  | 2015 |  |  |  |  |  |

**Table 12: The same data frame in Tidy Data Frame.**

*Table . Microsoft Excel object of the data-frame shown in Table 6*



# Part- 2 : Data and Analyses

The chart below compares the TOTAL\_MKT\_VALUE in 4 zones. For comparison, 1224 parcels chosen randomly from the Hamilton County, Cincinnati were compared against CORE, CENTER and EDGE. Source data for years 2007-2014 ( outlined in teal color) along with forecast data for years 2015-2018( outlined in red) is displayed in the plot below.

* Parcels in Buffer zones have had large TOTAL\_MKT\_VAL compared to the non-Buffer Zone.
* The CENTER zones exhibit largest increase in Total Market Value.
* Then it is closely followed by EDGE.
* The changes in 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 street car. Too close seems less desirable.

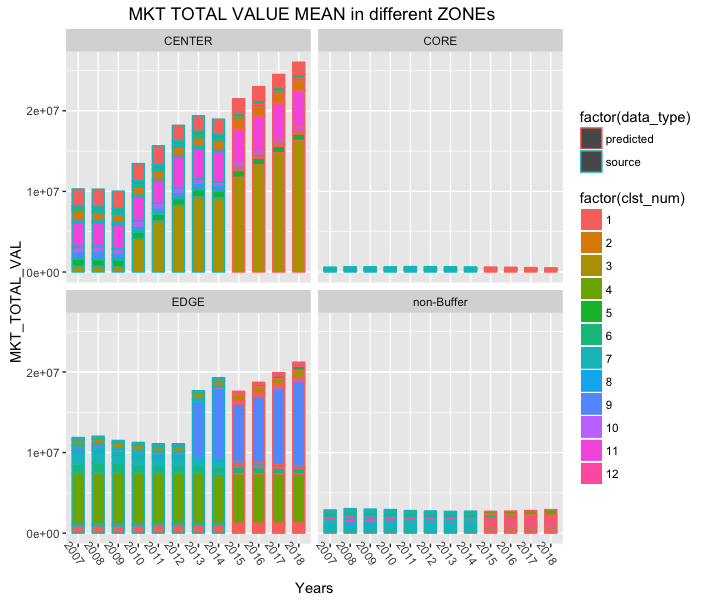


Figure : comparison of TOTAL\_MKT\_VAL in different zones

# Macintosh HD:Users:rajesh:Desktop:Coursera:SpringBoardGithub:Streetcar0719:Reports:Final Milestone:Plots:Compare_CENTER_EDGE.png

Figure :Comparison of CENTER and EDGE zones

A comparison between CENTER and EDGE zones, shows, that the forecast for TOTAL\_MKT\_VAL of the CENTER zone is more than the EDGE zones for the years 2016-2018. Within the CENTER zone, the properties within cluster 3 show most significant growth in TOTAL\_MKT\_VAL

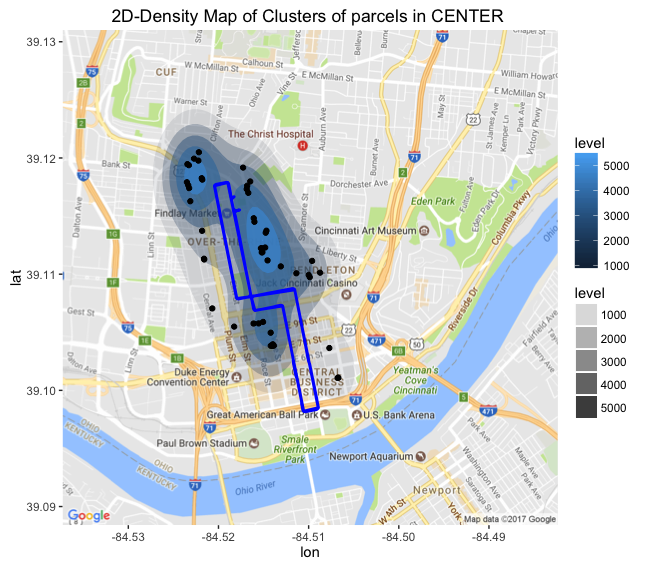
**rownames(partCENTER100$var[[3]])**

[1] "X0800002026300" "X0750001023000" "X0940004015500" "X0790004015400" "X0800002026800" "X0840002000100" "X0800002024300"

[8] "X0800002028400" "X0940007001400" "X0800002026500" "X0760001001000" "X0800002027100" "X0800002025900" "X0950002006800"

[15] "X0800002024400" "X0800002025000" "X0800001007300" "X0800002024200" "X0790004024600" "X0800002023600" "X0800002027300"

[22] "X0800002024900" "X0800002024600" "X0940005015200" "X0760002017700" "X0800002023300"

****

In the EDGE zone, the growth comes from Parcels in Cluster 9.

**rownames(partEDGE100$var[[9]])**

[1] "X1340005053000" "X0800001012100" "X0750002008700" "X0800001012600" "X0750002015700" "X1330006021500" "X0750002001500"

[8] "X0960004018700" "X0750001011600" "X0960004006300" "X0800001014500" "X0940007025490" "X0750002003600" "X0940005018300"

[15] "X1340003032900" "X0750003014200" "X0940007024300" "X0740001000100" "X0770001009500" "X1450001015900" "X0800001014600"

[22] "X0750003010900" "X0940007024400"

# Macintosh HD:Users:rajesh:Desktop:Coursera:SpringBoardGithub:Streetcar0719:Reports:Final Milestone:Plots:Compare_CORE_non_Buffer.png

Figure : Comparision of CORE and non-Buffer zones

The CORE buffer zone has fewer total observations (48) compared to 1224 observations in the non-Buffer zone. While, the two zones are not comparable, the non-Buffer zone trend shows declining TOTAL\_MKT\_VAL from the years 2008-2013 and then a forecast of increasing values for the years 2015-2018.

Within the CORE zone, a few parcels in Cluster 12 show declining trend.

**rownames(partCORE100$var[[12]])**

[1] "X1330003004500" "X1330003004100" "X1330003001600"

Within **the non-Buffer zone** Cluster 4 shows **increasing** trend

**rownames(part1500$var[[4]])**

[1] "X5980080028800" "X6030016010100" "X5930008018000" "X5100102089900" "X5920015005200" "X5000353010200" "X0300001007300"

[8] "X6030008008500" "X5900201073400" "X1940011021300" "X6120131009100" "X2140005000600" "X5900300005200" "X0810002033500"

[15] "X0080002016200" "X2100074011500" "X5100101004600" "X5100102121000" "X5100102038100" "X6120040039400" "X2120068000200"

[22] "X5000270026500" "X1270002005900" "X6510005013900" "X0420A01005900" "X5400112008000" "X0340002000900" "X5940010043500"

[29] "X0750002005500" "X2240004007000" "X2080060010900" "X0670002006500" "X5200240006100" "X5610009017900" "X0230001005400"

[36] "X5500081046300" "X5100251003300" "X6000080023300" "X5400070023000" "X5610009010800" "X5900162011600" "X0660001001000"

[43] "X5700020001100" "X6000211031800" "X6110011019300" "X6000202017300" "X6080008018700" "X1170009010600" "X5980060029900"

[50] "X5250014002700" "X0390007002900" "X6200060017500" "X0280005008600" "X5100102108800" "X5000220033100" "X5260130003200"

[57] "X5500170024100" "X5100051000200"

And in the same non-Buffer Zone Cluster 5 show **declining trends** :

**rownames(part1500$var[[5]])**

[1] "X5400032030300" "X2080060011900" "X0500002020100" "X5260050004700" "X6000091021500" "X1820005000200" "X5500170006100"

[8] "X5500154001100" "X2080067018900" "X5000290035700" "X0260002008600" "X5500161019900" "X2240006009500" "X5000133003800"

[15] "X1240002005200" "X0390004013900" "X1720018003400" "X6110020031300" "X6210003009700" "X0390002031300" "X5000351019200"

[22] "X6510017000700" "X1890022007900" "X5000341040200" "X5280001013600" "X5000481009200" "X6000011034100" "X0390002025100"

[29] "X5000072003300" "X0500007022200" "X0980002008600" "X0570006005100" "X5000281020100" "X5000201007700" "X6080015010400"

[36] "X5200173013100" "X1830002037900" "X5270021005000" "X5210009008100" "X0490003009000" "X1950031005700" "X0390002018600"

[43] "X0590006010500" "X5990050055500" "X0860001006100" "X5400011000400" "X5000341036400" "X1780028012600" "X1300002007900"

[50] "X0020007011100" "X5100052059000" "X5000171009700" "X6510056005800" "X5910012029700" "X2480001015000" "X5500081008100"

# Part-2: Methodology

## Extraction Transformation and Loading

**Step1:** 9 years of data from the was bound into a single data frame df\_all. The original data included 13 predictors on 2,191,115 observations

Call :

df\_all <- bind\_rows(df\_taxinfo\_2007, df\_taxinfo\_2008, df\_taxinfo\_2009,df\_taxinfo\_2010,df\_taxinfo\_2011,df\_taxinfo\_2012,df\_taxinfo\_2013,df\_taxinfo\_2014)

dim(df\_all)

[1] 2191115 13

**Step2:** To scope down the project, it was decided to focus on TOTAL\_MKT\_VAL predictor.

df\_MKT\_TOTAL\_VAL <- df\_complete\_cases %>% select(one\_of(c("DF\_TAXINFO\_YEAR","PARCEL\_ID", "MKT\_TOTAL\_VAL")))

**Step3:** 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 of data was used for study

**Call : taxinfo\_matrix <- df\_MKT\_TOTAL\_VAL %>% spread("PARCEL\_ID", value="MKT\_TOTAL\_VAL", fill=as.numeric(0))**

**dim(taxinfo.matrix)**

**[1] 8 241748**

## Standardization

**Step4**: A **time series object** and analysis for every 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 (mean sample size of the 3 buffer zone) parcels picked at random from a population of 241,748 parcels. This represents about 0.51699 % of the population size.

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

**taxinfo.matrix\_scl <-scale(taxinfo.matrix\_uniq, scale=TRUE, center=TRUE)**

## Clustering

**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. 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 being the sum of the squared correlation between the variables and the center of the cluster, which is the first principal component of PCAmix.

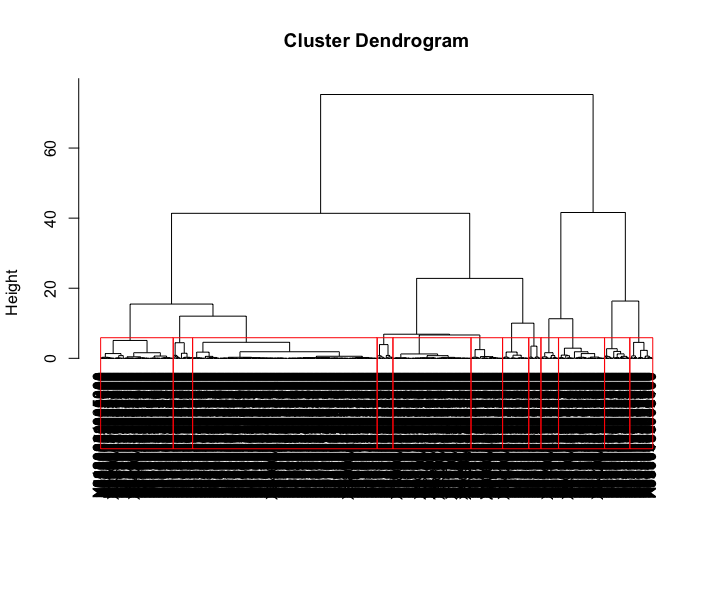


Figure : Dendogram of non-Buffer zone of 1224 sample parcel ids from the Hamilton County

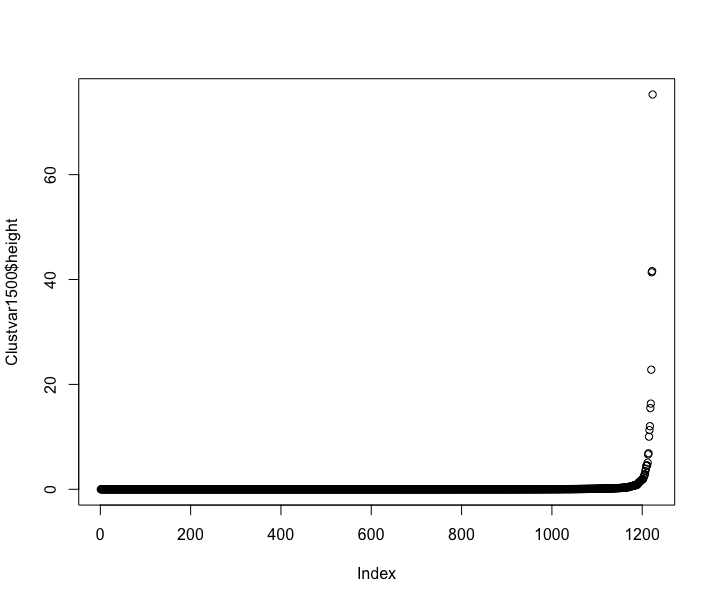


Figure : Based on intuition and the plot of height variable of the hclust object, a cluster of 12 seemed reasonable choice for further analyses

A cluster size of k = 12 was chosen.

## Forecast Model Development and Selection

### **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.

A user defined function call on 12 clusters, plots 4 parameters to check for forecast errors

Call :

**# Use Holt-Winters Simple Exponential Smoothing**

**for ( clst in c(seq(1:12))) {**

**forecast\_clusters(clst,a, taxinfo.matrix\_uniq)**

**}**

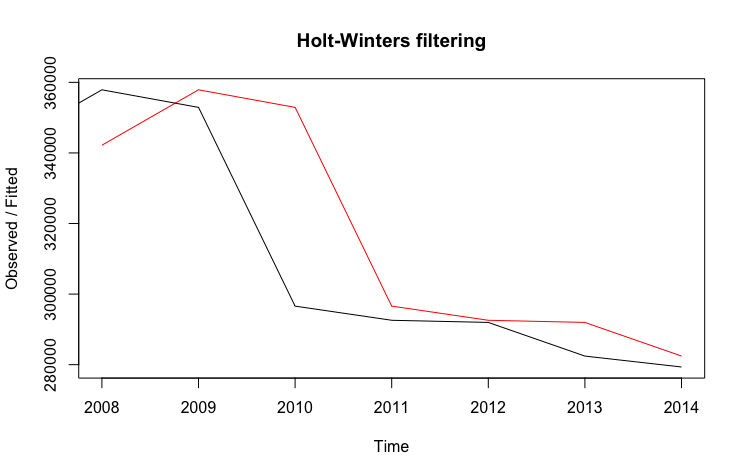


Figure : Actual vs. Fitted values of in-sample source data. The black line shows the observed in-sample source data for the years 2007 – 2014. The red line marks the fitted data i.e. the prediction of the forecast model.

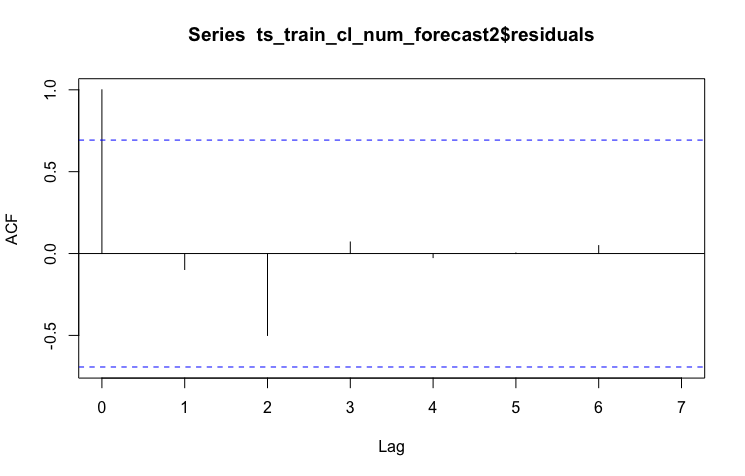


Figure : Autocorrelogram indicates that the forecast errors lie within significance bounds. This was the observation for all of the clusters.

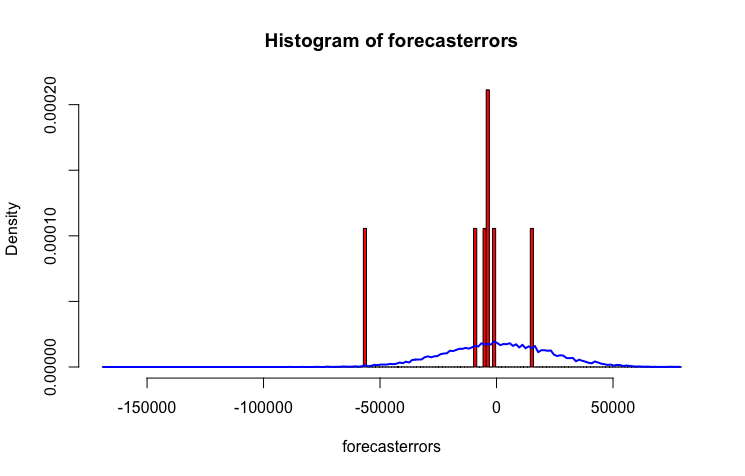
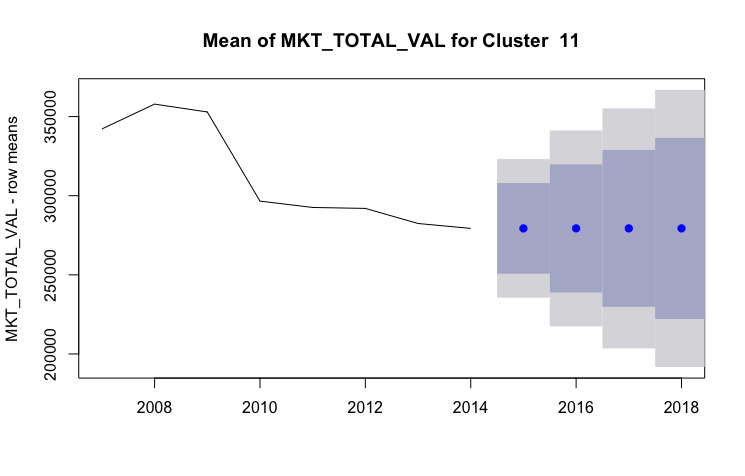


Figure 21: Left : Holts Winters Forecast Source data and Forecast data plotted with the 80% and 95% confidence intervals. Right Histogram of forecast errors shows that the variance is not close to 0

In Fig. 19, even though the ACF lies within the bounds of significance levels, it is observed it is observed in Fig. 21, that the variance of forecast errors is not zero and therefore, the Holts Winters Exponential method was rejected.

### **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, 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. The actual vs. predicted for each cluster and for each polynomial order was plotted. And a Mean Absolute Percentage Error was computed for the test set.

# 

Figure : Overview of Methodology: Data was extracted and loaded. It was filtered and re-shaped for time series forecast, models were created, model accuracy was measured to select accurate models for prediction and source and forecast data were compared with Buffer zones( CORE, CENTER and EDGE) against other zone

Using a dendogram, it was decided to cut the tree into 12 clusters. Intuition played a role in the selection of the number of clusters.

To make forecasts using simple exponential smoothing in R, we can fit a simple exponential smoothing predictive model using the “HoltWinters()” function in R. To use HoltWinters() for simple exponential smoothing, we need to set the parameters beta=FALSE and gamma=FALSE in the HoltWinters() function (the beta and gamma parameters are used for Holt’s exponential smoothing, or Holt-Winters exponential smoothing, as described below). **HoltsWinters()** was applied to rowMeans() of each cluster.

Using the “forecast” library, the predictive model fitted using the HoltWinters() in the variable was passed to **forecast.HoltWinters()** function for predicting the Annual Taxes for next 20 years.

## **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 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 a **problem** with the data :

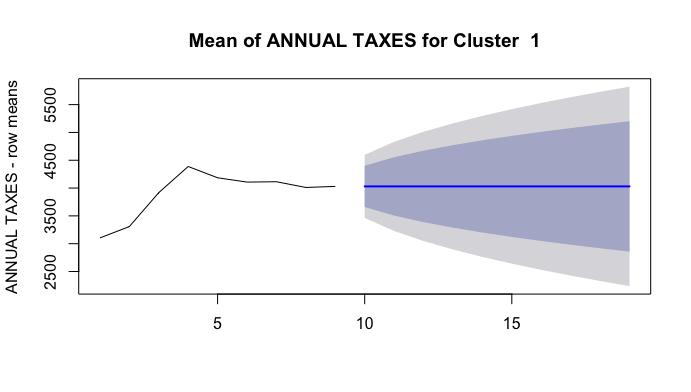


Figure : HoltsWinters Simple Exponential smoothing and prediction for cluster 1. The source data mean for years 8 and 9 indicating of problem data set.

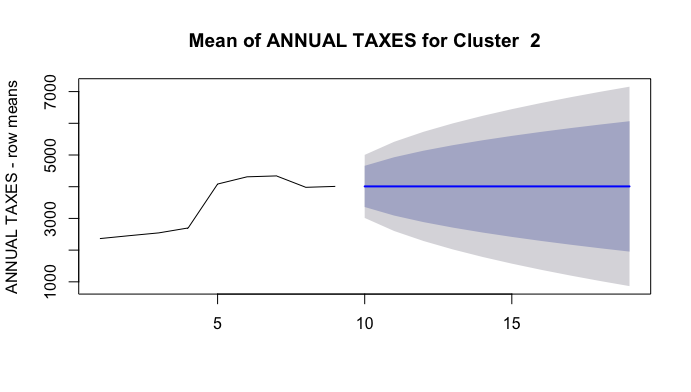


Figure : HoltsWinters Simple Exponential smoothing and prediction for cluster 2. The source data mean for years 8 and 9 indicating of problem data set.

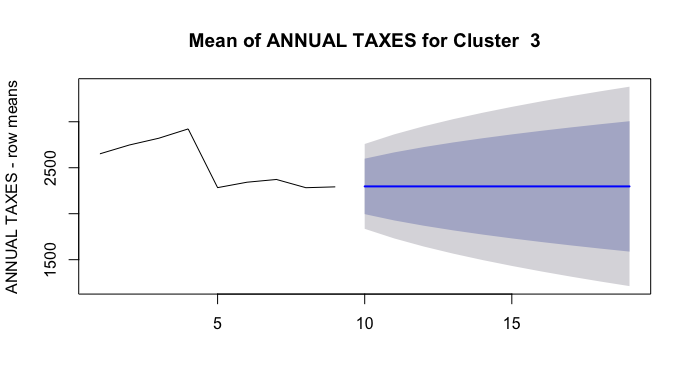


Figure : HoltsWinters Simple Exponential smoothing and prediction for cluster 3. The source data mean for years 8 and 9 indicating of problem data set.

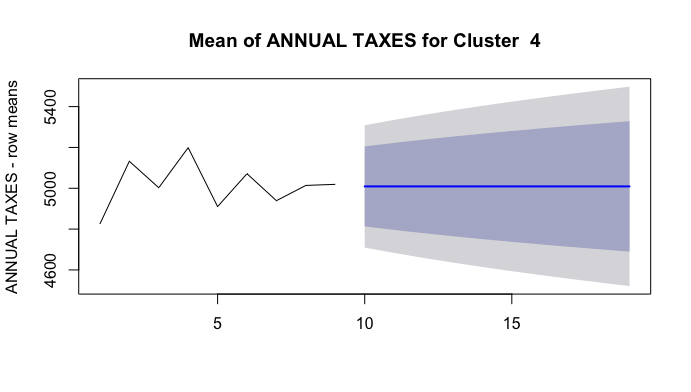
****

Figure : 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 :** Also, it was decided to perform the study on Total Market Value 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. Also, it was decided to use data for 8 years instead of 9 years.

**Problem :** Using HoltWinters on 9 years of annual data, one could not

**Solution :** It was decided to

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/>