K. Rajesh Jagannath

Mentor : Anirban Ghosh

Forecasting TOTAL MARKET VALUE of Parcels

Introduction of a Street Car in Cincinnati

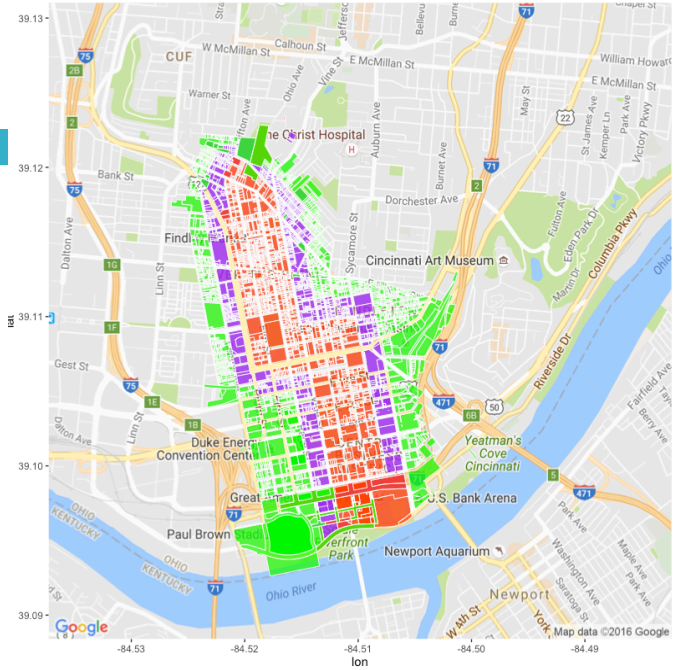
Motivation

- In August of 2016, a streetcar was introduced in the downtown Cincinnati, OH.
- The introduction of the Street car has an Economic impact
- 70,000 people work in the downtown

- It provides easy access to business, work and home
- It draws new business, permit fees, and **property tax**
- It is disruptive to the neighborhood (crowding, transient population, noise)
- QUESTION was – What is the impact on economic indicators ? negative or positive -

Introduction

- A measure of economic impact is the Annual Taxes assessed on property parcels
- Annual Taxes= f(TOTAL_MKT_VAL)
- OBJECTIVE: Predict the MKT_TOTAL_VAL for 2015-2018
- SCOPE : A Buffer Zone was established spatially around the Street car route
  - CORE
  - CENTER
  - EDGE



Data Sources and Data Challenges

Data sources were

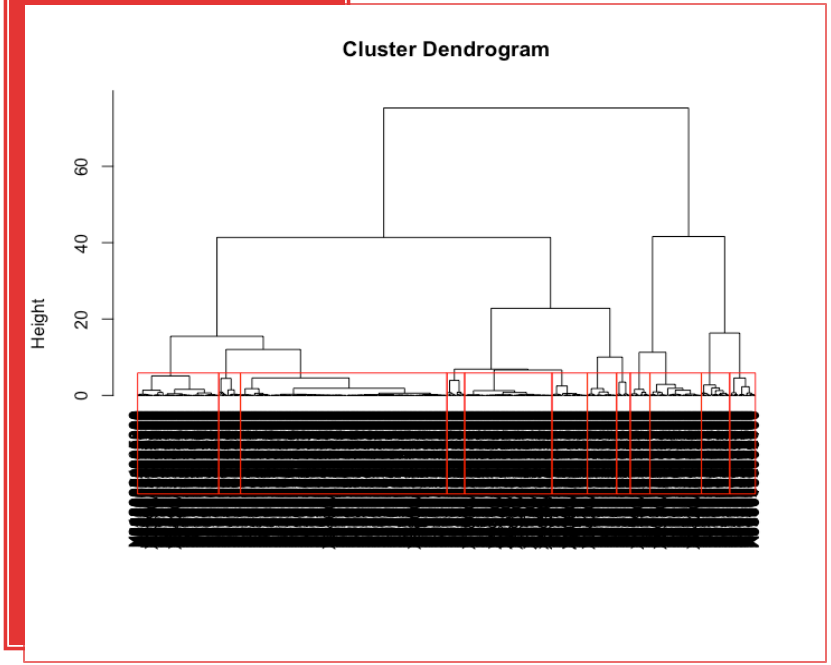
- Cincinnati Area Geographic Information Systems (CAGIS), City of Cincinnati, OH and
- The Hamilton County Auditor Office, Cincinnati, OH

- 9 years Annual data on all the parcels in the Hamilton county ~ 290,000 observations/year
- Data was extracted, transformed and loaded to form a data frame 2,190,994 x 13 predictors
- MKT_TOTAL_VAL was the feature selected for study
- Cleaning of data, fixed width format, and the sheer number of observations was a challange

MKT_LAND_VAL
MKT_IMPR_VAL
ANNUAL_TAX
EXLU_CODE
TAXES_PAID
DELQ_TAXES
FORECL_FLAG
ACRE
SALE_AMT
SALE_DATE
MKT_TOTAL_VAL
VALID_SALE
NEW_COSNSTR

Clustering of Parcels

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



- The data was reshaped into 2,190,994 X 3
- forecast on every parcel would have been time consuming and resource intensive
  - A random sample representative of the number of parcels CORE, CENTER, BUFFER Zone
- Further, using ClusterOfVars library, strongly related parcel ids were clustered together
- By plotting the dendogram as well as intuition, it was decided to cluster each zone into k=12 clusters

Prediction : polynomial regression

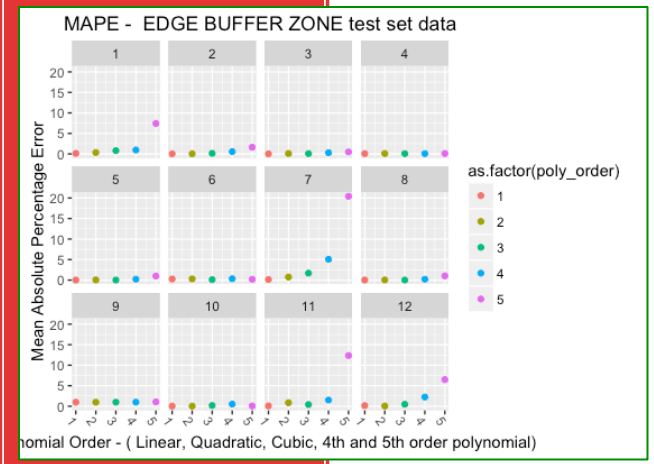
Polynomial Regression

- Next, Polynomial non – linear regression was chosen for trend forecast of MKT_TOTAL_VAL
  - Linear, Quadratic, Cubic, 4th and 5th order of polynomial models were considered
  - SOURCE DATA
    - Years 2007-2012 – TRAINING
    - Years 2013-2014 – TEST SET
  - PREDICTED DATA
    - Years 2015-2018 - PREDICTION
  - HoltsWinters Time Series was initially assessed
    - due to variance of the forecast errors it was discarded

Accuracy of selected models

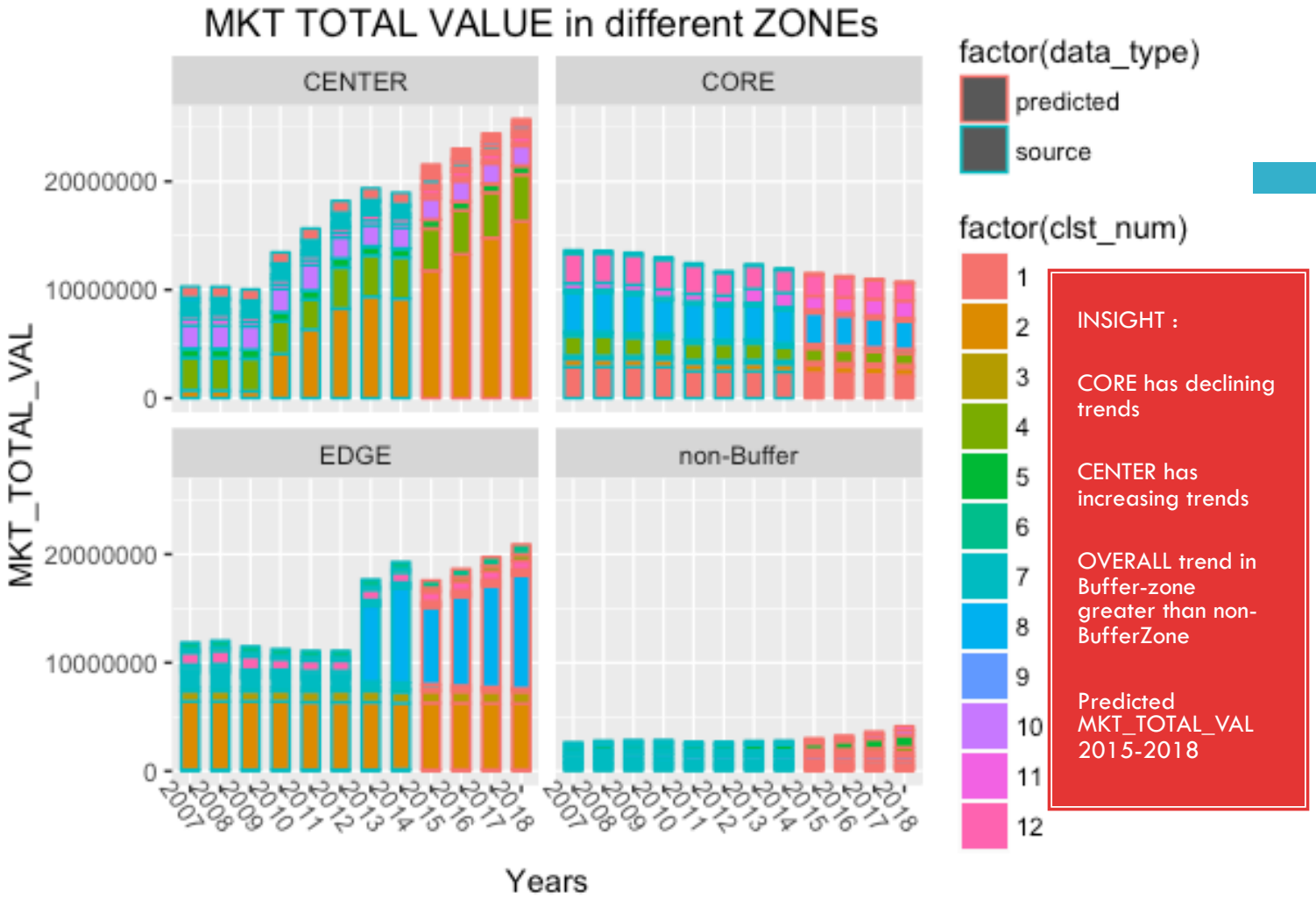
Mean Absolute Percentage Error (MAPE)

$$\left(\frac{1}{n}\sum\frac{|Actual - Forecast|}{|Actual|}\right)*100$$



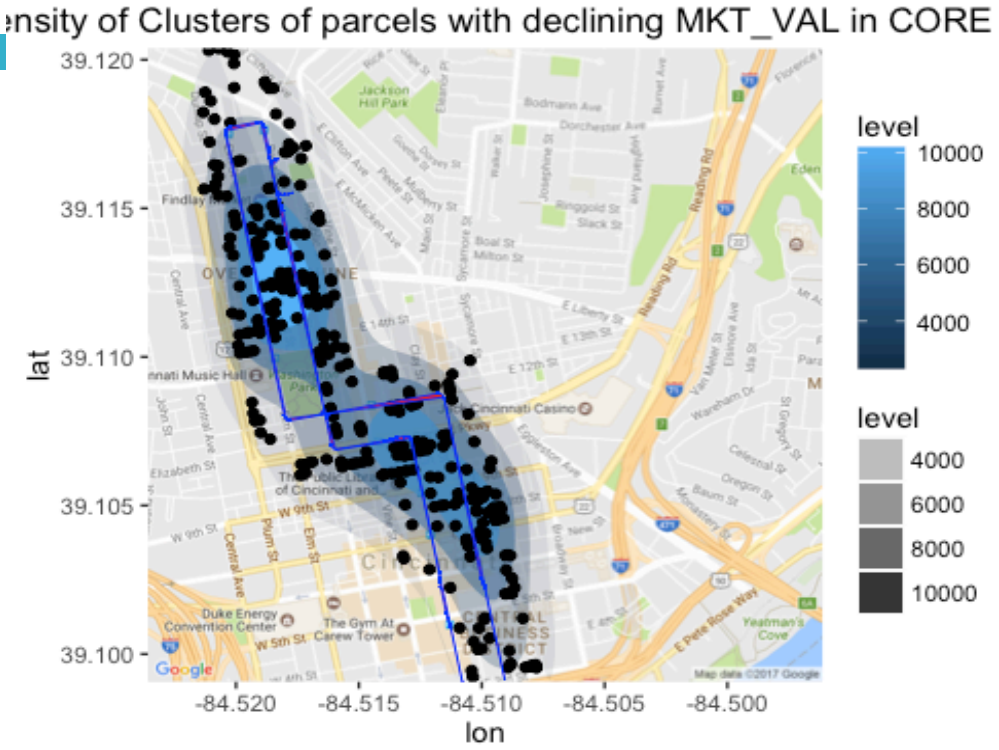
- For Each Zone in {CORE, CENTER, EDGE, non-Buffer}
  - For each cluster in {1 ..12}
  - Mean Absolute Percentage Error(MAPE) was computed for polynomial order {1,2,3,4,5} polynomial regression model
  - Models with least MAPE were selected to perform prediction
  - Lower order models had Error within 5%

Vizualization



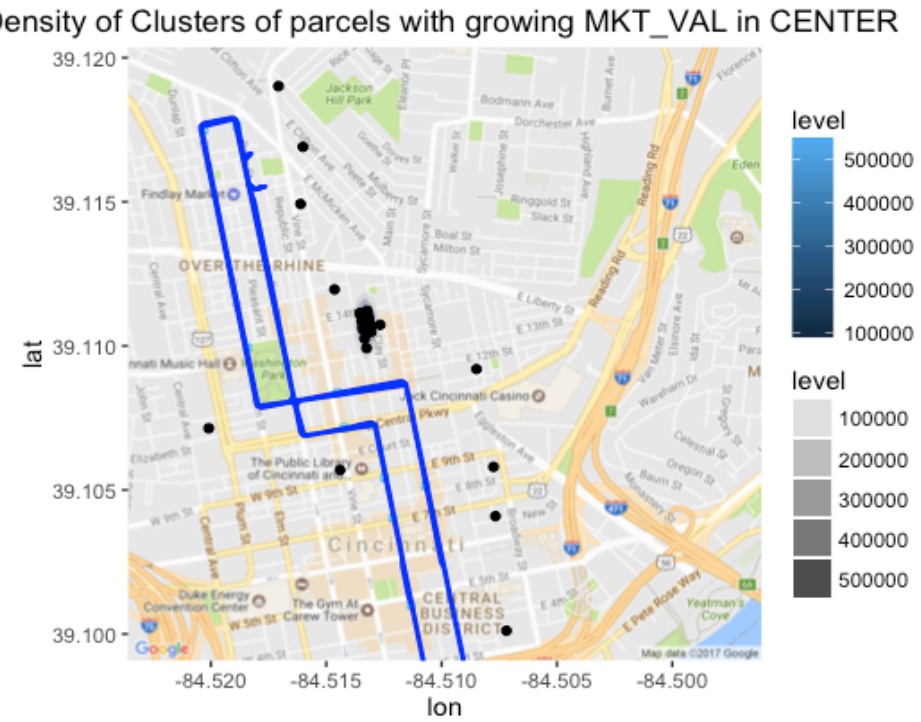
Core Zone

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 denser concentration of parcels bounded by E. Liberty St. and E. 14th St. and between Elm St. and Race St.



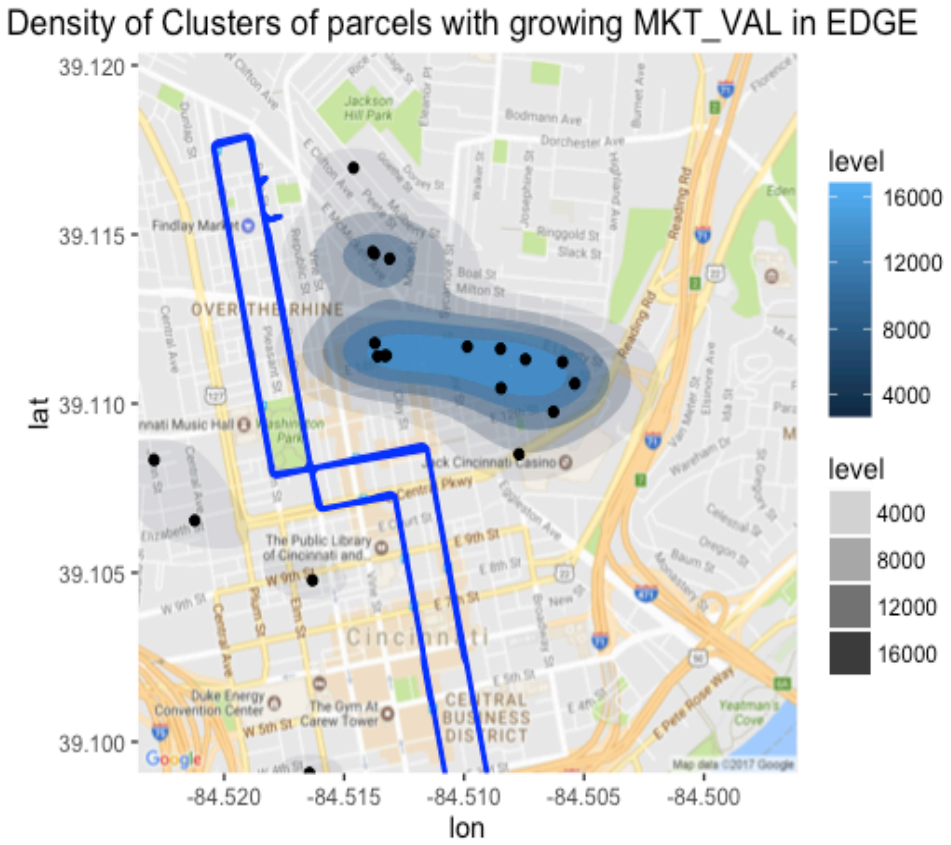
CENTER Zone

26 parcels of cluster 2 in the CENTER dominate increasing trends. There was a 38.78% increase of MKT_TOTAL_VAL from $305M(2015) to $424M(2018). The spatial distribution appears densely co-located around Walnut St and Clay St. and E 14th and E 13th St.



EDGE Zone

23 parcels of cluster 8 in the EDGE Zone dominate increasing trends. It seems that there is a 46% increase of MKT_TOTAL_VAL from $163M (2015) to $239M (2018)



Conclusions

ETL

Clustering

Training

Testing

Prediction

Visualizations

- For Descriptive Analysis, data was Extracted, Translated and Loaded in to data frames : A majority of the time was spent on getting clean data
- The data from CAGIS was taken for 9 years (2007-2015) and filtered and reshaped into data frames with MKT_TOTAL_VAL selected as the primary predictor
- For each buffer-zone, parcels were clustered into 12 clusters
- For prediction, non-linear polynomial regression model was created for each of the 12 clusters using years 2007-2012 as training set
- Accuracy of the model was calculated using the Mean Absolute Percentage on a test set for years 2013-2014. Most accurate models were selected.
- Clusters of parcels with increasing and decreasing MKT_TOTAL_VAL trends in each of the areas under study were visualized on a ggplot with Google maps as one of the layer

Problems Encountered

Problems Encountered	Solution
Data was not clean i.e not from a competition or academic data set	Needed insights and clean-up from staff at CAGIS
Using on-line geo-coding was not reliable to get Longitude and Latitude	Requested CAGIS to provide the location data
Data was in 3 distinct fixed width format	Different scripts were used to read in Data
Many observations for Test data for Year 2015 seemed exactly the same as observations in test data for 2014	Data for 2015 was not used for testing the model instead test-data was for Years 2013-2014
Performing time series on every parcel required time and number crunching	Used Sampling, Dimension reduction, Clustering and limited the scope for convergence
I was unprepared for the time it took to clean the raw data	Reduced the scope to analysis on MKT_TOTAL_VAL

Acknowledgments

People

and

Resources

- Mr. Anirban Ghosh (Springboard Mentor)
- Mr. Raj Chundur (Director, CAGIS)
- Resources
  - http://www.cincinnati.com/story/news/2016/05/05/streetcar-nation-kc-opens-friday-cincy-next/83874740/
  - https://cran.r-project.org/web/packages/ClustOfVar/ClustOfVar.pdf
  - http://www.exegetic.biz/blog/2013/12/contour-and-density-layers-with-ggmap/
  - http://www.shanelynn.ie/massive-geocoding-with-r-and-google-maps/
  - http://stat405.had.co.nz/ggmap.pdf
  - A little book of R for Time-series, *Avril Coghlan*
  - An Introduction to Statistical Learning, *Gareth James et al*

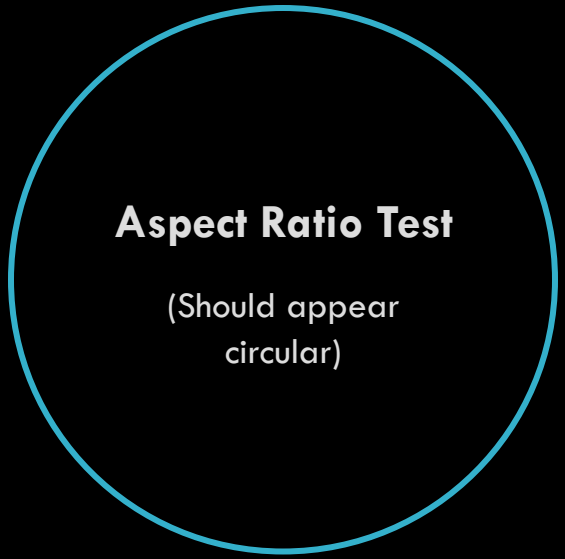
Future Work

Other Predictors

More recent data

- Analysis can be performed on other predictors such as business permits, bldg improvement permits, restaurant permits
- The Analysis has been performed using the years the Street car was not yet in service
- More recent data from Auditors office improve the accuracy of the forecast

Widescreen Test Pattern (16:9)



4x3

16x9