STA 208 Project Report - NYC Taxi Data Pickup Prediction

Xingtai Li, Guicheng Wu, Wenhao Wu

University of California, Davis

gchwu, xtali, wnhwu@ucdavis.edu

June 2, 2016

Content

Background

- Objective: predicting the number of taxi pickups within a certain period of time and a neighborhood in New York City.
- Potential Applications: more efficient cab dispatch, trip planning, traffic monitoring, etc.
- Related projects:
  1. MIT Big Data Challenge, 2013.
  2. Todd W. Schneider, "Analyzing 1.1 Billion NYC Taxi and Uber Trips, with a Vengeance," 2015.
  3. Di-Tech Challenge, 2016.

Data Sets Processing Techniques

Data Sets

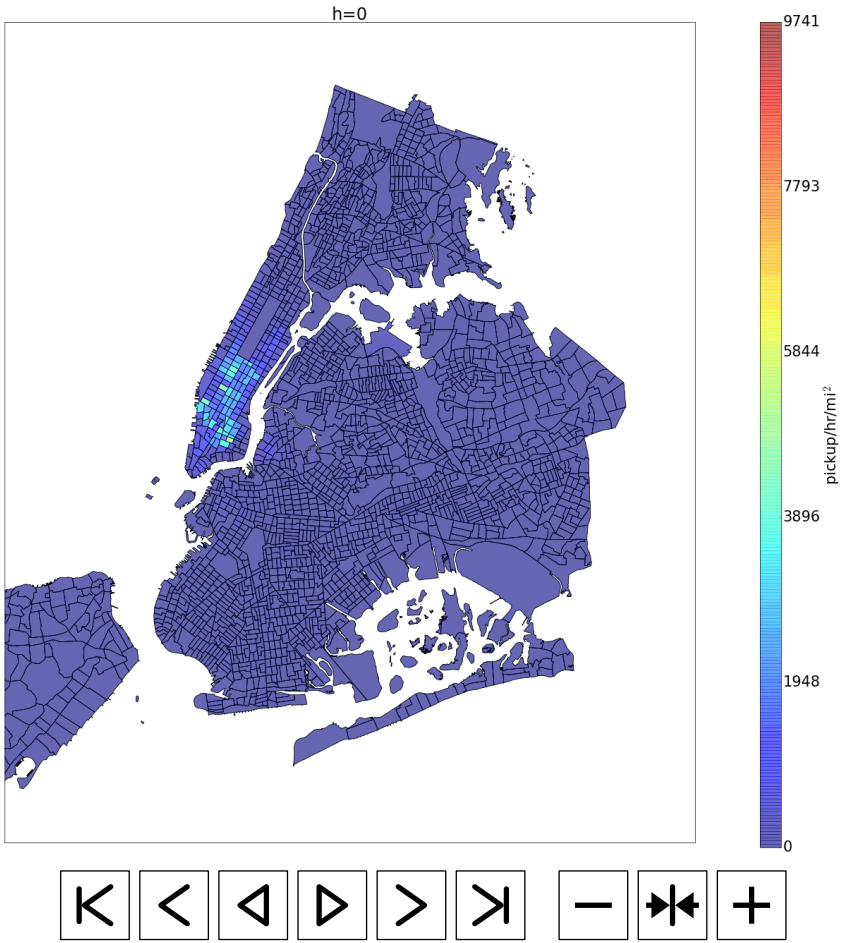
- NYC TLC Taxi Data Year 2014 (25GB), 177312293 complete records.
- NOAA daily average temperature and precipitation measured by New York Central Park Belvedere Tower observing station.

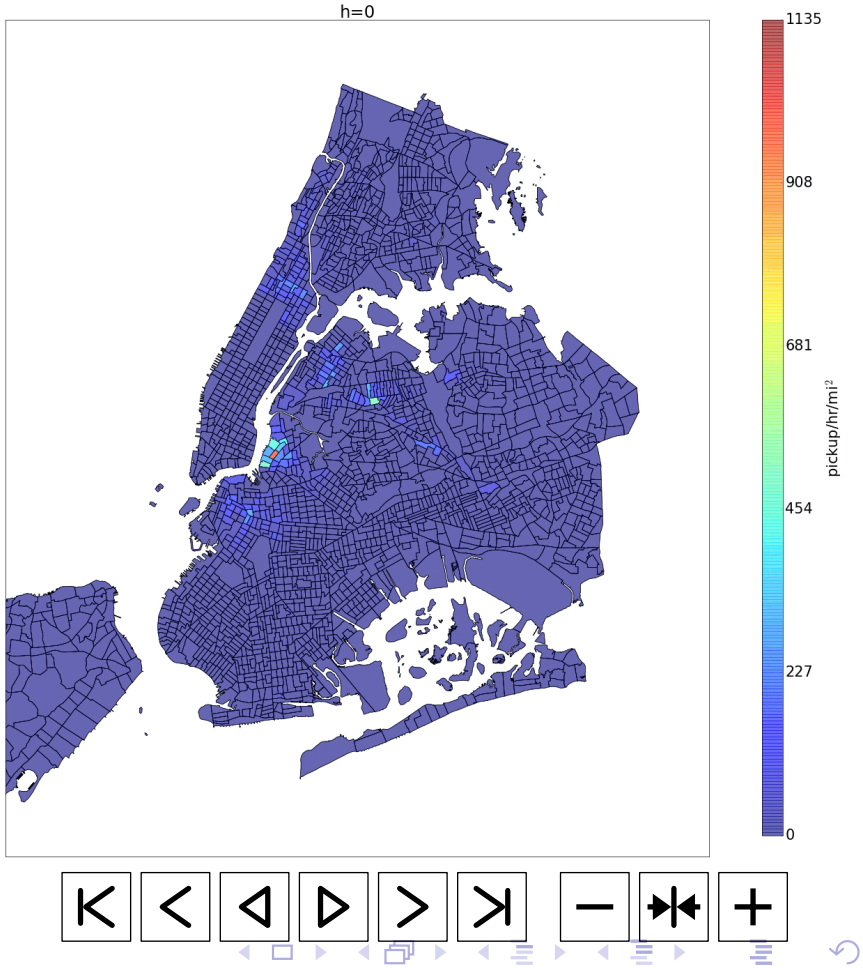
Processing Techniques

- Data import: PostgreSQL+PostGIS via SQLAlchemy+GeoAlchemy.
- Machine learning: scikit-learn.
- Visualization: Matplotlib+Basemap.

Demo: NYC TLC Taxi Data

Pickups/hr/mi^2 in 24 hours (averaged over 365 days)





Demo: NYC TLC Taxi Data

Time series analysis: hourly number of pickups (365×24)

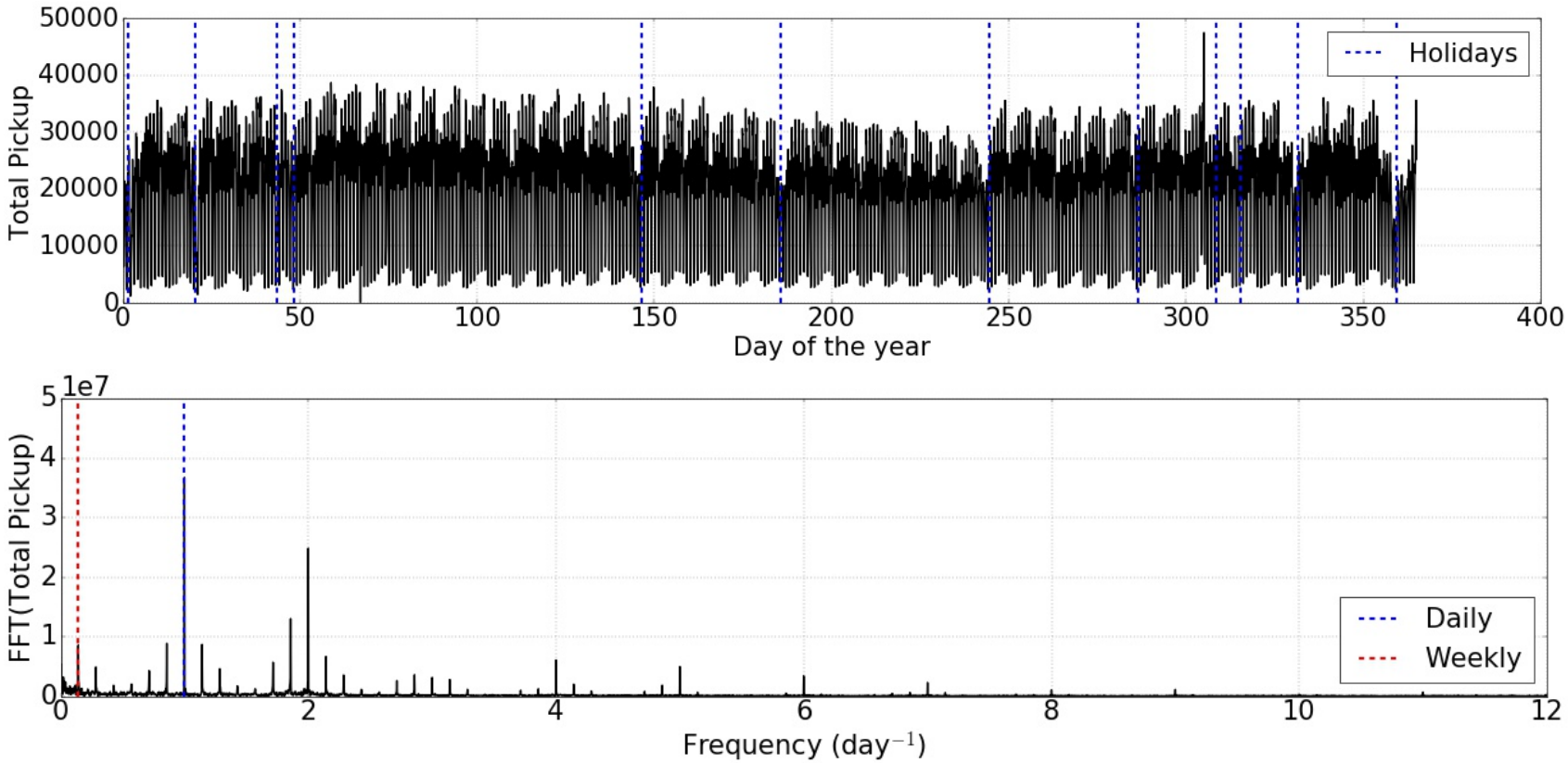


Figure : Temporal and frequency pattern of the total pickup count.

Feature Selection

Table : Feature selection.

Feature name	Data Type	Meaning
pickup gid	categorical	The id of pickup neighborhood.
pickup dow	categorical	Day of week (0..6)
pickup hour	categorical	Hour of pickup (0..23)
temperature	numerical	Daily Temperature
precipitation	numerical	Daily Precipitation
holiday	categorical	True/False
count_1	numerical	pickups at previous hour

Categorical features are encoded $\longrightarrow$ a total of 35 features (7 (DOW)+24 (hr)+temperature+precipitation+holiday+count_1).

Machine Learning Models

1. Linear regression.
2. Lasso.
3. Support ridge regression.
4. Bayesian linear regression.
5. Decision tree.

Performance Metrics

- For each neighborhood (NTA) a regression model is trained independent of other NTAs.
- Performance metric: R^2 coefficient of determination

$$R^2(y, \hat{y}) = 1 - \frac{\sum_{i=0}^{n}(y_i - \hat{y}_i)}{\sum_{i=0}^{n}(y_i - \overline{y})}, \overline{y} = \frac{1}{n}\sum_{i=0}^{n} y_i \quad (1)$$

as well as mean squared error (MSE)

Model Comparison (1/3)

- Training/testing set split: 80% vs 20%.
- Only the 26 NTAs with the average number of pickups per hour per square mile greater than 100 are studied.

Table : Training and testing results of the 5 regression models for Upper East Side-Carnegie Hill

Regressor	Parameter	R^2	RMSE
Linear	-	0.8987	128.5
Lasso	$\alpha = 0.01$	0.9005	130.51
SVR	rbf kernel, $C = 10^5$	0.6710	241.61
Bayesian Ridge	$\alpha = 10^{-5}$,$\lambda = 10^{-6}$	0.9009	131.06
Decision Tree	max length =100,min leaf = 10	0.9241	104.8

Model Comparison (2/3)

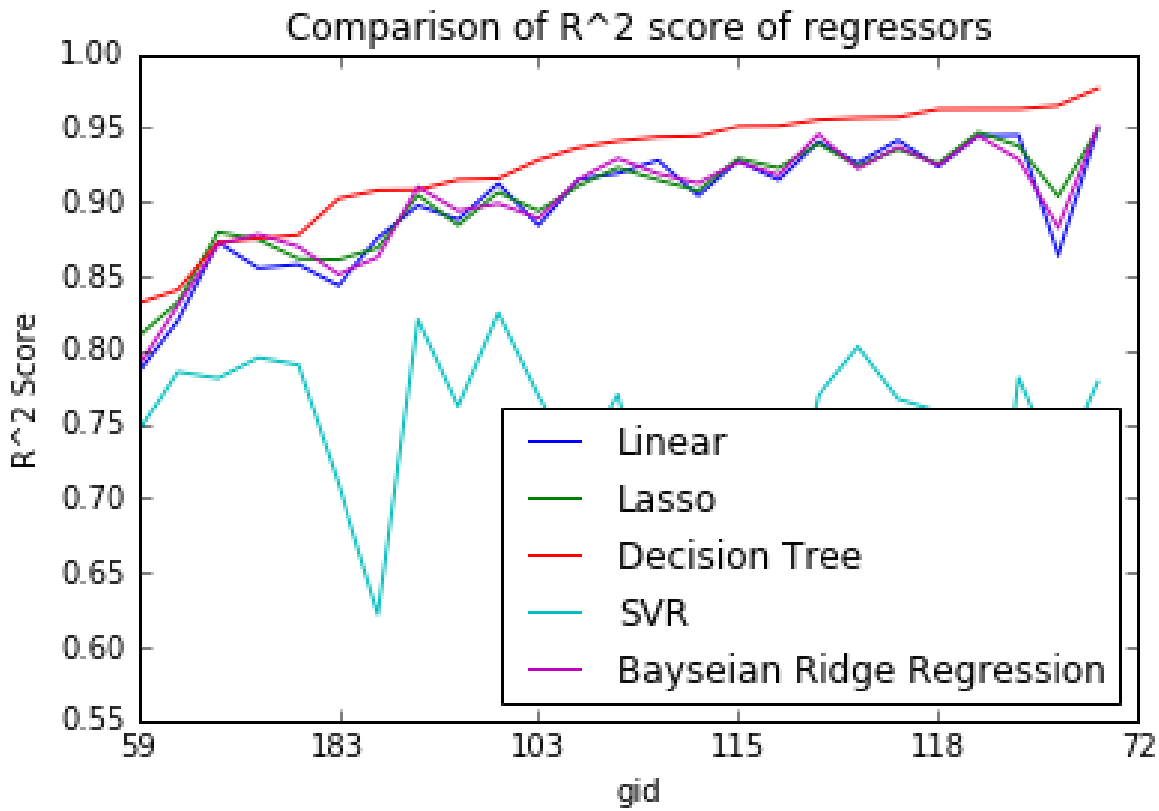


Figure : Comparison of the regressors in R^2 scores for all 26 selected NTAs.

Model Comparison (3/3)

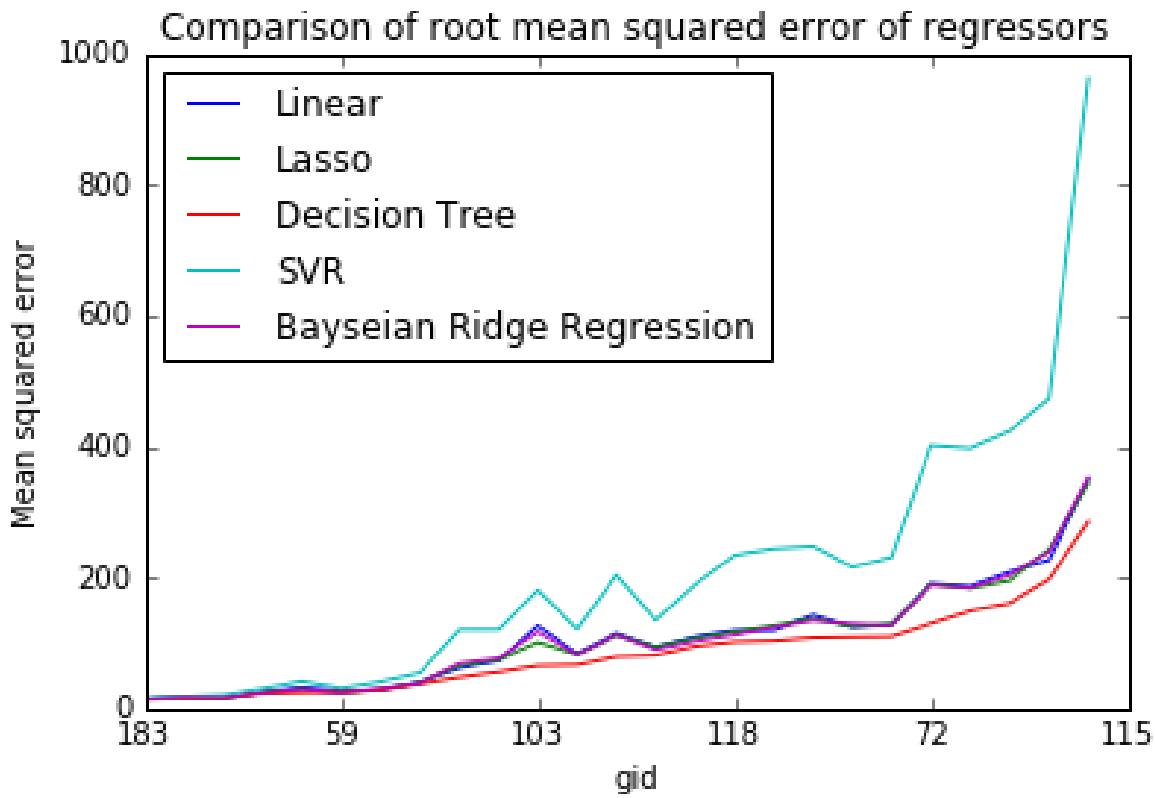


Figure : Comparison of the regressors in MSE for all 26 selected NTAs.

Training Curve

Focusing on the decision tree for Upper East Side-Carnegie Hill (gid=27).

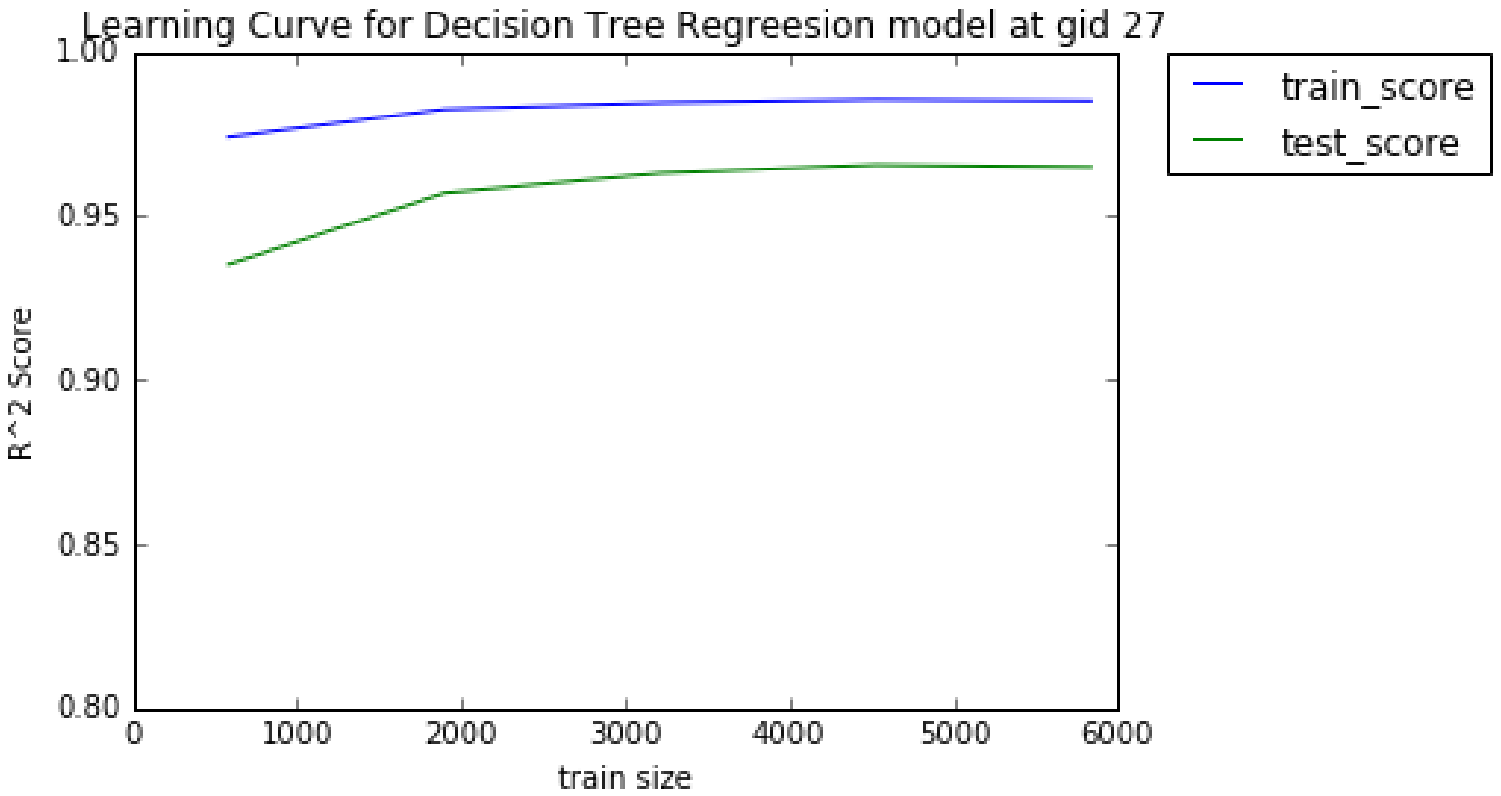


Figure : Learning curve for the decision tree.

Weights of Features

Table : Weights of features in the decision tree

DOW	0-6	0.0167
Hour	0-23	0.1076
Temperature	numerical	0.01098
Precipitation	numerical	0.00256
Holiday	True/False	9.91e-05
count_1	Numerical	0.86196

Decision Process

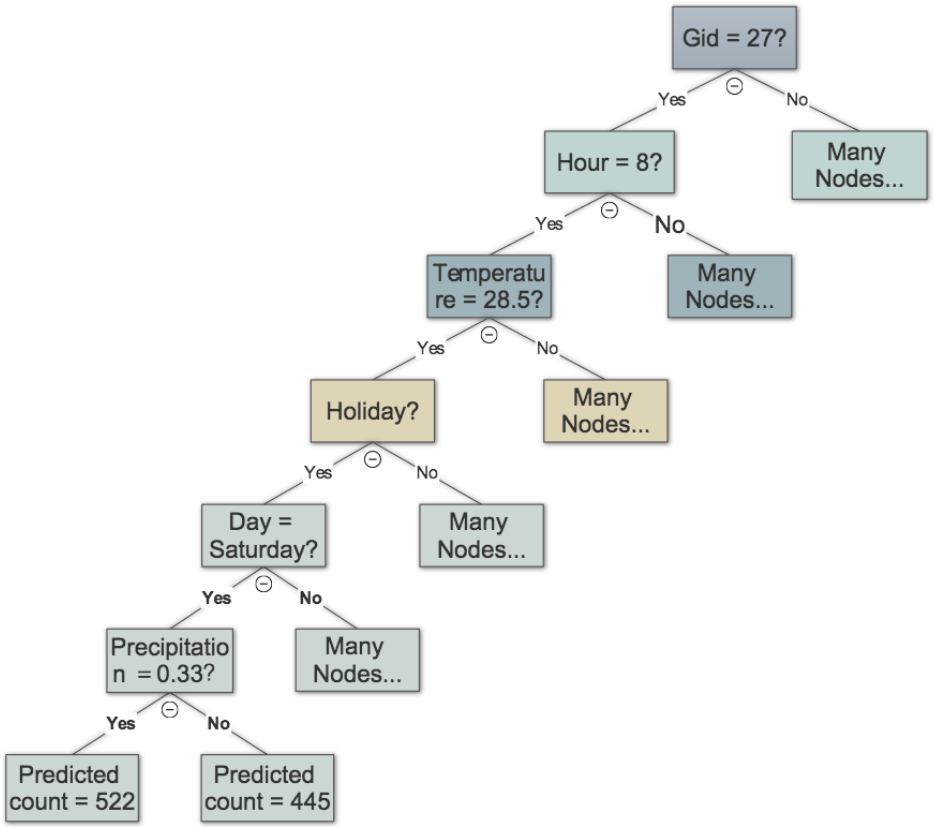


Figure : Final trained decision tree

Prediction vs True Value

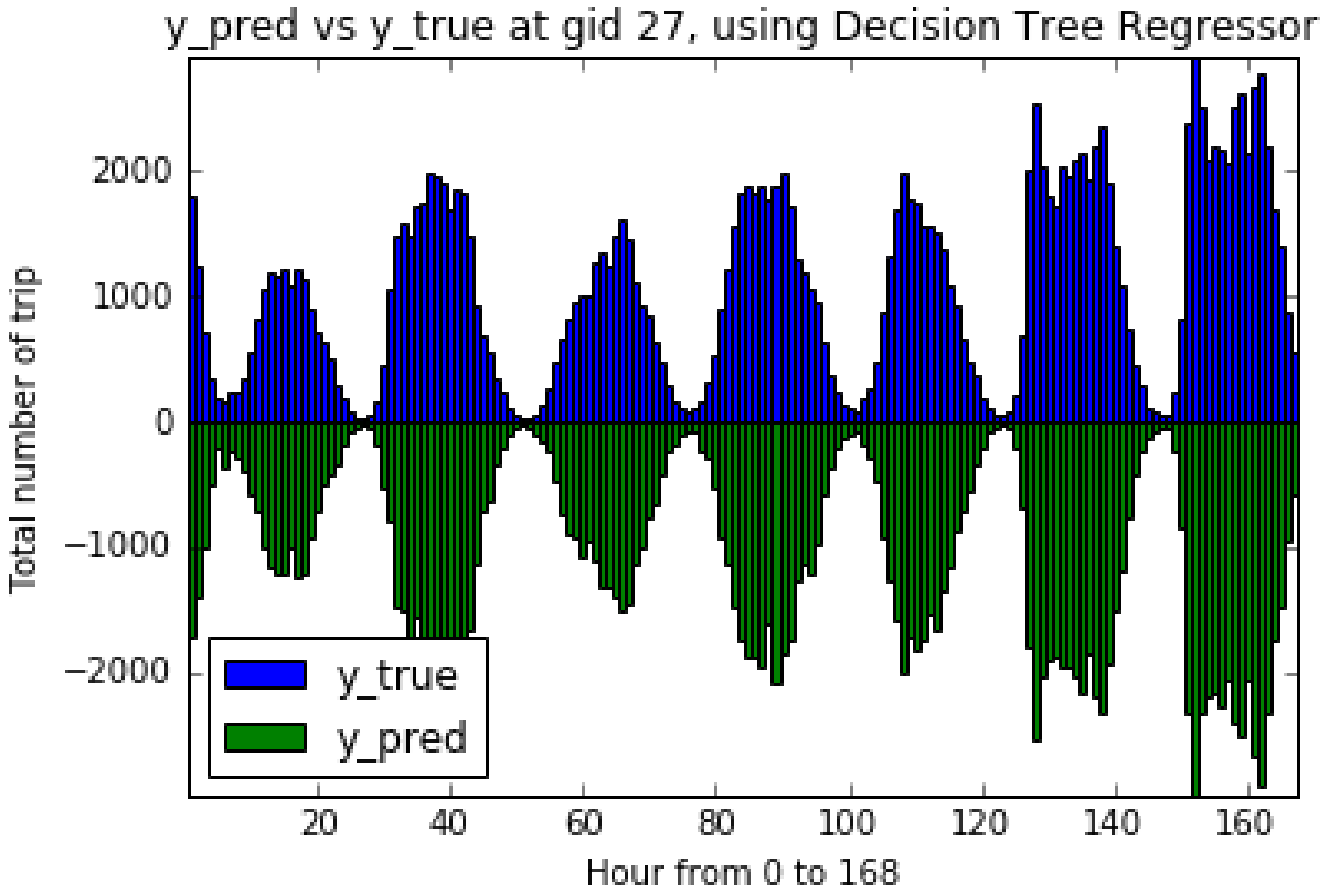
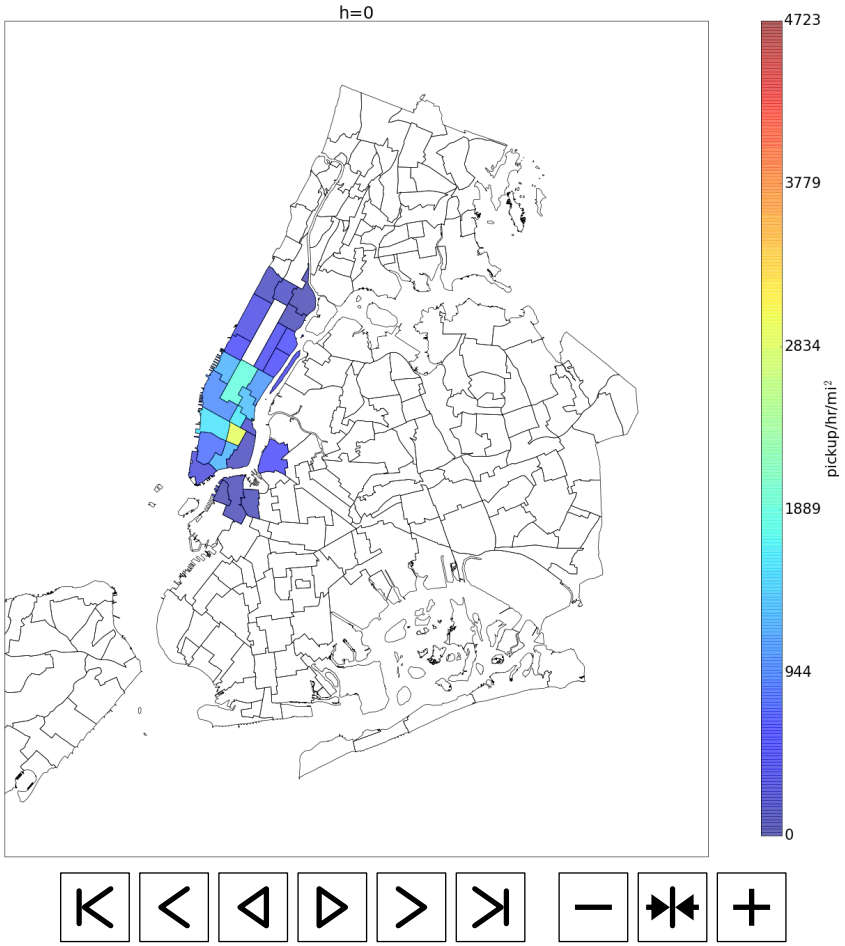


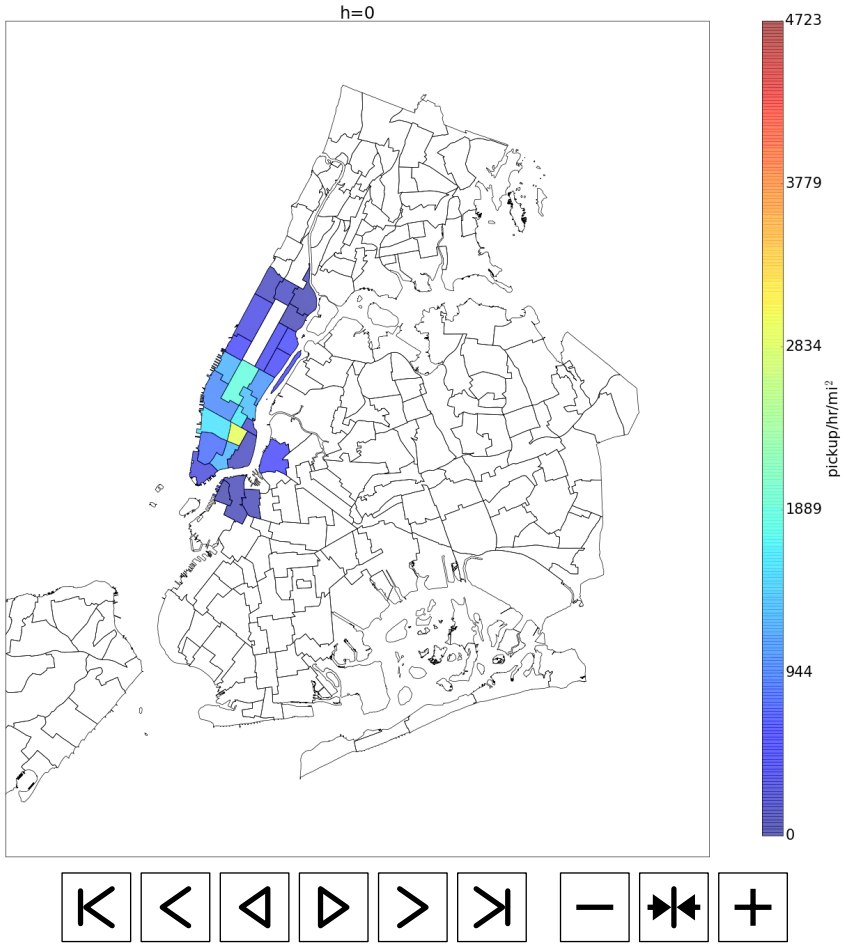
Figure : The predicted hourly count of trips vs the true one.

01/20/2014, Monday, Martin Luther King Jr. Day

Predicted

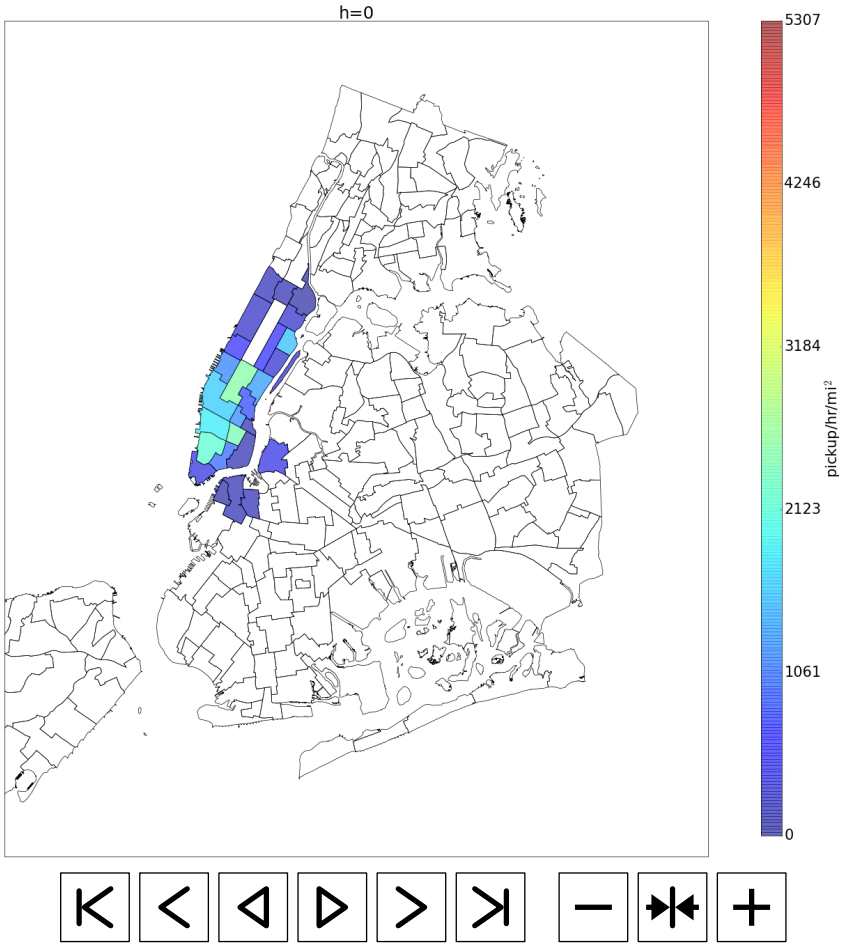


True

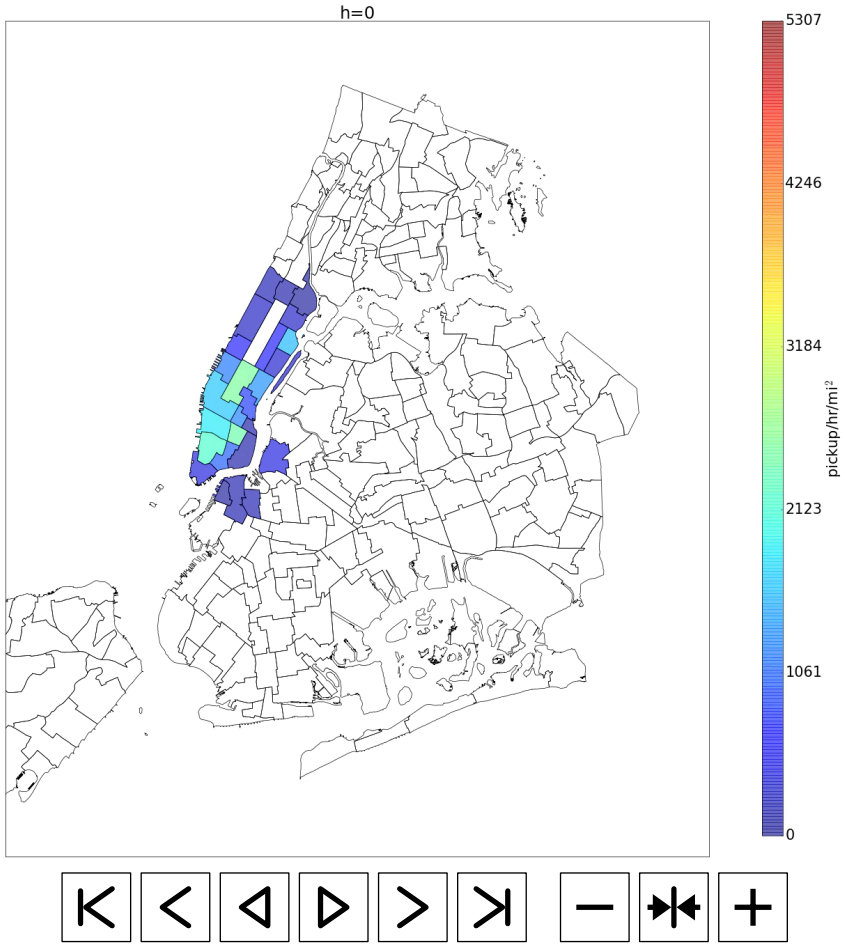


04/10/2014, Thursday

Predicted

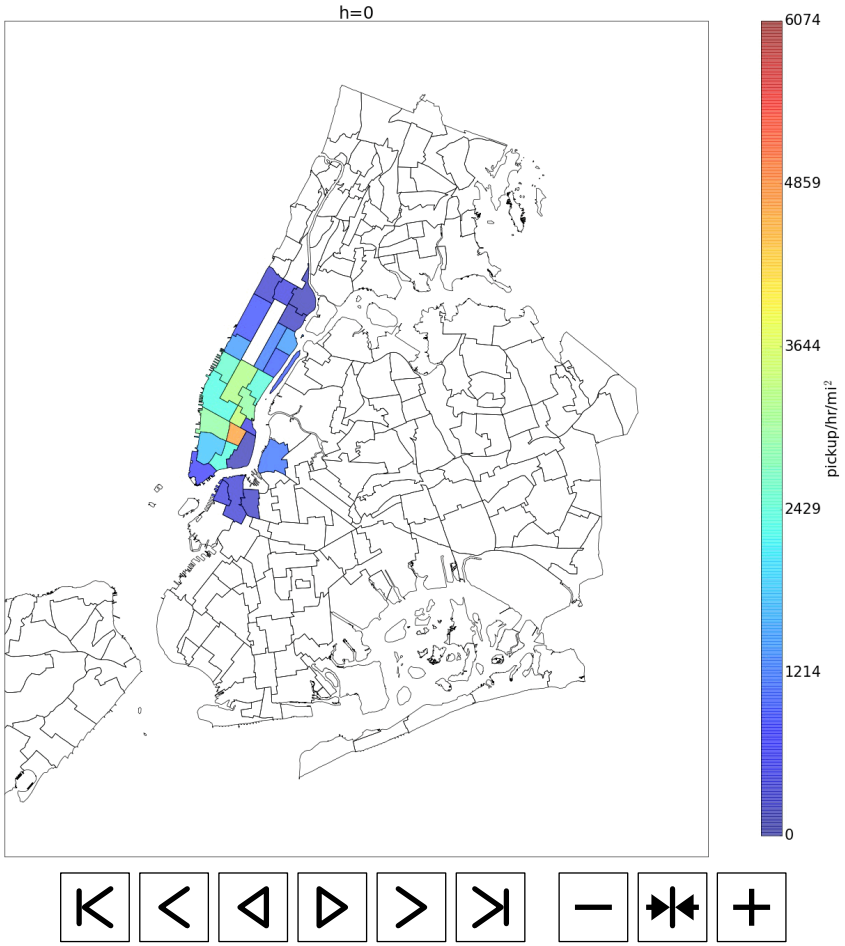


True

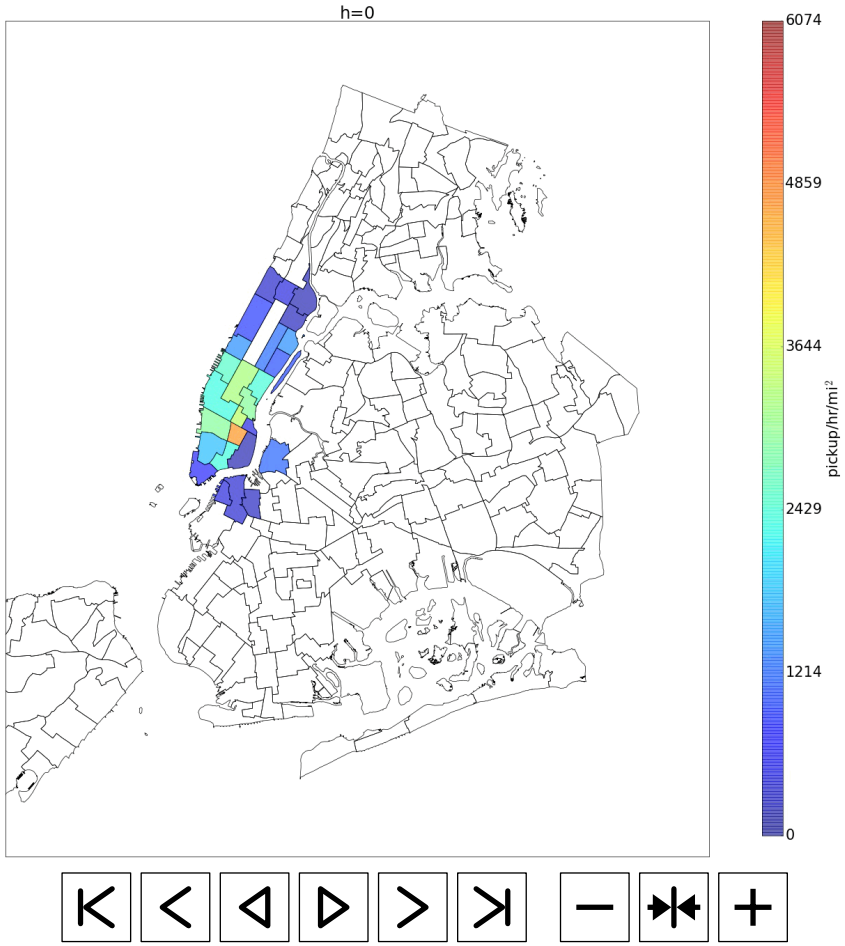


11/09/2014, Sunday

Predicted



True



Conclusion

Contribution

- A rather accurate and practical model to predict the NYC taxi pickup built atop of various raw data sets.

Future Works

- Finer spatial and temporal resolutions.
- Extend the dataset across different years.
- Including more features: longer history, traffic, buildings in each neighborhood, etc.
- Explore other information: dropoff distribution, fees, etc.