Hurricane Sandy: 311 open data analysis of pre and post

**Final Paper: Foundation Module**

# TEAM MEMBER

1. Kim, Tae ([thk301@nyu.edu](mailto:thk301@nyu.edu))
2. Putro, Dimas Rinarso ([drp354@nyu.edu](mailto:drp354@nyu.edu))
3. Zhuang, Yuzheng ([yz2611@nyu.edu](mailto:yz2611@nyu.edu))

# Introduction

New York City has its own story on how it bounced back from one of the biggest disasters ever hit the East Coast since the 20th Century, Hurricane Sandy. With 147 direct deaths reported, in which 72 of m occurred in the mid-Atlantic and Northeastern US, Hurricane Sandy caused the biggest direct fatalities related to tropical cyclones outside the southern areas since 1972 (Eric et al., 2013). This assignment paper examines 311 dataset changes and patterns in the year 2011, 2012, and 2013 respective to the time range close to Hurricane Sandy (October 1st – December 31st), to see unseen patterns that can be extracted. This paper focuses into determining impact of Hurricane Sandy by first analyzing trend changes in the year before, during and after Sandy in relation to 311 top complaints data against agencies. Then, the paper describes series of applicable statistical models and tools that we have learned from this course that were being used.

# Software and tools

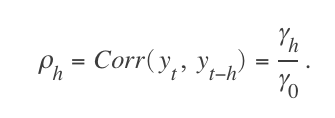
All data cleaning, logarithmic calculation, re-indexing and visualization processes were conducted using Pandas library in Python. Minor manual adjustments were applied in Microsoft Excel. Data cleaning processes comprises trimming and indexing the data, temporarily writing Pandas dataframe into intermediary format such as CSV so that all team members can use it in other programs. Visual representations such as map were generated using Bokeh library.

# Time series analysis of complaints against specific agency

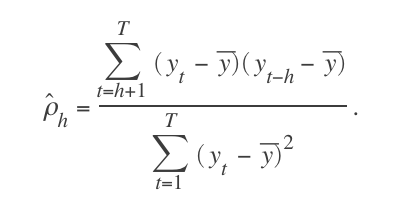
Our first step is to categorize the complaints by agency and plot the time series to find visually see the patterns. From figure 1 to figure 3 we could see that regardless of the scale of number of complaints, NYPD and TLC were showing minimum response to the event of disaster, constantly maintained the trend without noticeable peaks. DPR, on the other hand showed peaks in 2011 and 2012 in figure 1 and figure 2 respectively, with in 2012 it reached maximum value of 4400. This provided us hints that DPR trends in 2011 and 2012 differ compared to 2013. Another clue provided by this time series plot is that the negatives trends NYPD complaints in 2012 when DPR reached it peak value. From this we understood that we wanted to limit our scope of observation to only DPR and the next step to statistically confirm the state of abnormalities for DPR.

# Detecting Abnormalities: ACF and PACF

Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) was preferred to auto regression method since we wanted to particularly set a threshold test to see if in specific year DPR could be justified to have abnormal behavior, specified by the level of randomness between data point in certain time interval. This could be achieved by observing the relation of the data at *t* and *t*-*h*, where *h* is a time lag and in this case was set in the range of 1-20 days, and then see whether it falls within certain threshold of level of confidence (95%):



The denominator γ0 is the lag 0 covariance. The autocorrelation function (ACF) for a time series *yt*, t = 1,...,N, is the sequence *ρh, h* = 1, 2,...,N – 1. The partial autocorrelation function (PACF) is the sequence ϕh,h*,h* = 1, 2, ..., N – 1.The lag-*h* autocorrelation would be then obtained by :



To visually observe this, we could see the correlogram presented in figure 4 and figure 5 for ACF and PACF respectively. We could see from ACF (left) figure that in 2011 and 2012, the lag equal to 0 and 1 fell outside the confidence level threshold, while in 2013 lag = 1 was within the boundaries. This tells us that in 2013, the autocorrelation value for DPR fluctuates in smaller periodic in respect to time intervals. This was caused by increased randomness driven by number of complaints scattered. We also figured out that PACF, while showing similar pattern difference 2011-2012 vs 2013, was more difficult to observe see, thus, for setting up threshold of abnormality for natural disaster events, ACF was considered to be simple, relevant model complementing the use of manual time series observation.

# The use of maps and visual representations: what went wrong?

Since we know that statistically there were different in patterns of DPR’s number of complaints in 2011-2012 versus 2013, we could not tell what is the particular reason why the pattern was different and what variables were correlated. Thus, we did visual representations by plotting the number of DPR complaints per zip code by longitude and latitude given in 311 dataset, and plot it on the map. The result can be seen in figure 5, 6 and 7. From here we could observe that while Sandy’s effect is apparent in 2012, there were significant influx of complaints in Staten Island on 2011, while in 2013 is mostly flat. This might be related to the snow storm that in vast majority occurred in Staten Island in October 2011 (Silive, 2011). The next step is to find good predictor for regression analysis and build the model.

# Bivariate regression analysis

First, we obtained the number of trees data in New York City and try to find the correlation by running bivariate regression model with both linear and second order polynomial regression. This is to confirm that the 2 disasters, snowstorm in 2011 and Sandy in 2012, were assumed largely affected by trees. Our finding shows that in 2011 and 2012, in figure 8 and 9, characterized by increased R–squared value at 0.59 and 0.81 for both linear and second order polynomial fit, respectively. This indicates that the model explains most of the variability of the sample data around its mean. Linear fit, represented by black bold line, pointed an increased value from ß = 1.30 to ß = 1.83 from 2011 to 2012, while it decreased to to from ß = 0.77 in 2013, when there was no disaster events. 2011-2012 presented indication of super linearity correlation, meaning that the number of DPR complaints has more increased rate relative to the number of trees. However, 95% confidence level, as shown in figure 11-13 fell between 0.932 - 1.670 and 1.543 - 2.126 in 2011 and 2012 respectively, which tells us in 2011 there under the 95% level of confidence there was a possibility of sub linear characteristics (ß <1). The other interesting finding is that the both of the model, linear and second order polynomial could maintain the similar result, confirming that tree and the number of complaints do scale linearly and therefore no need to use more sophicasted model. This resonates with some emphasis given in class that “in most cases linear approach is the best model to start with”, as linear regression approach was also an interesting subject explored by other urban planners such as in identify the famous “economic of scale” concept proposed in (Luis and Geoffrey, 2007).

# Adding predictors and multivariate analysis

Our first step is to get better-quantified numbers of the neighborhood changes and

# Result and conclusions

Our first step is to get better-quantified numbers of the neighborhood changes and

# Table of figure

# Macintosh HD:Users:dimasrinarso:Dropbox:5004:tae_ads_foundation:2011_3agencies.png

Figure 2011 time series plot of top 3 complaints to agency in New York City

# Macintosh HD:Users:dimasrinarso:Dropbox:5004:tae_ads_foundation:2012_3agencies.png

Figure 2012 time series plot of top 3 complaints to agency in New York City

# Macintosh HD:Users:dimasrinarso:Dropbox:5004:tae_ads_foundation:2013_3agencies.png

Figure 2013 time series plot of top 3 complaints to agency in New York City

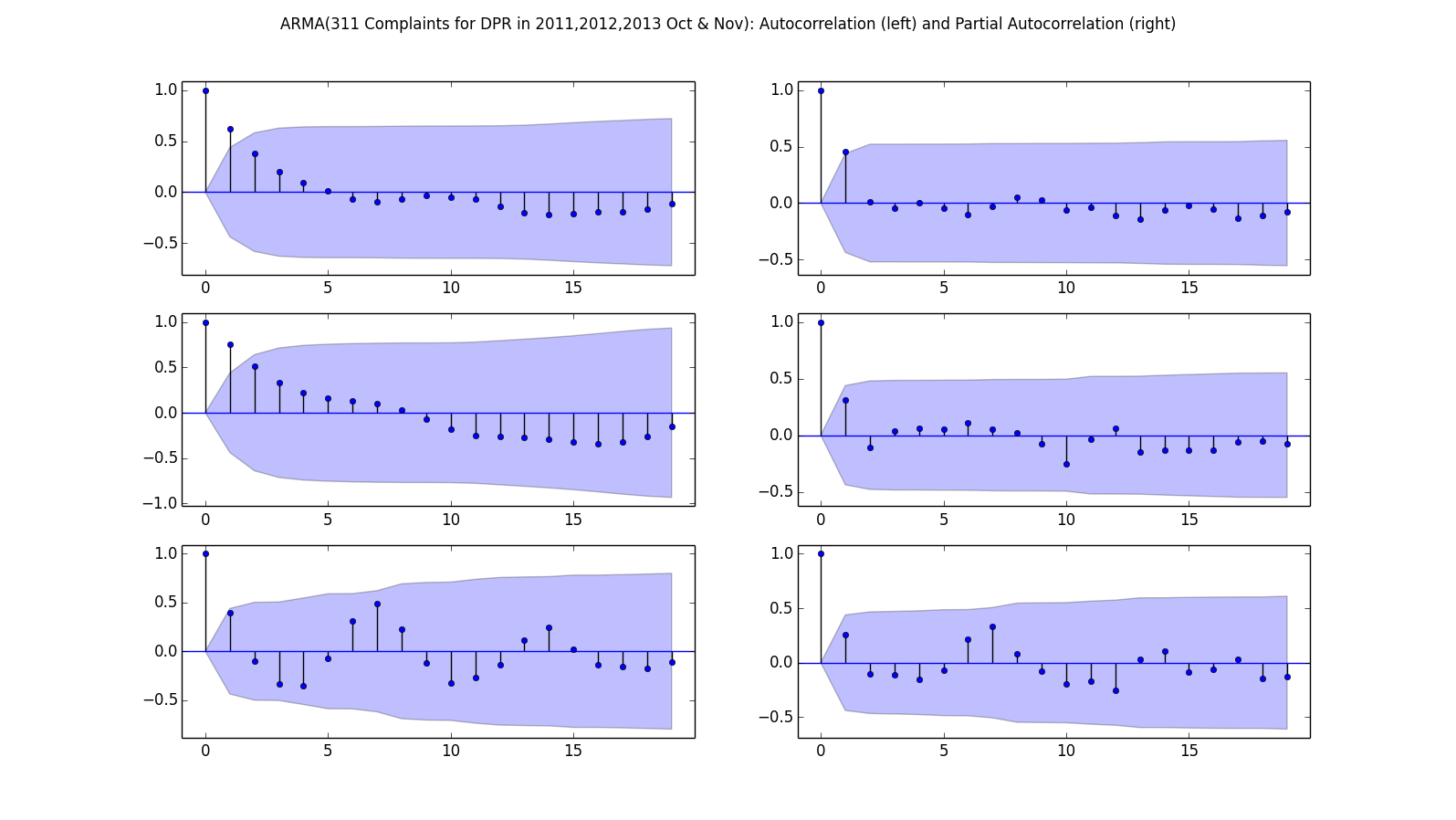


Figure ACF (left) and PACF (Right) of the DPR complaints for Oct and Nov 2011 (top), 2012 (middle), 2013 (bottom)

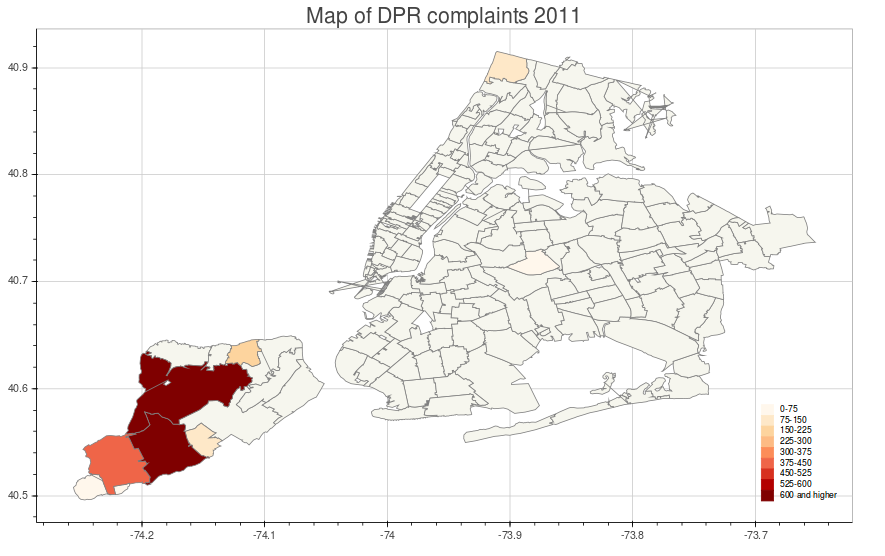


Figure Map of DPR complaints 2011

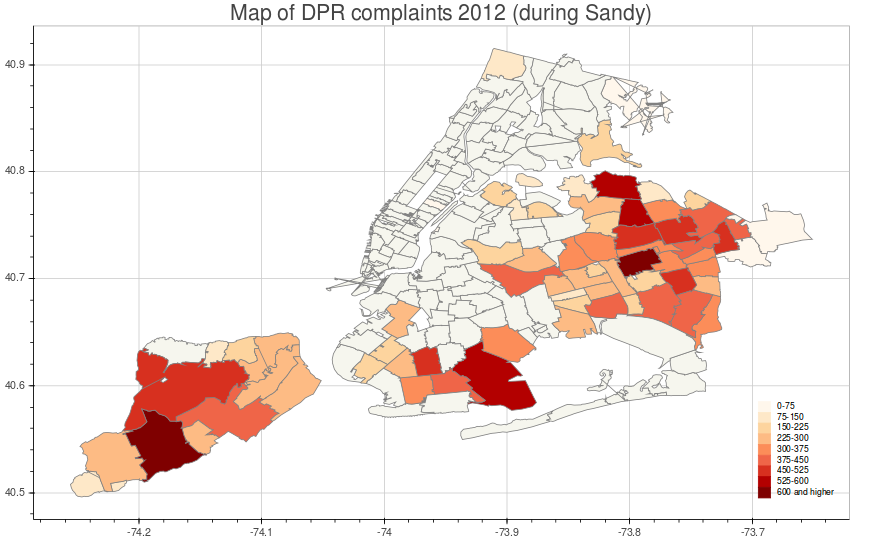


Figure Map of DPR complaints 2012

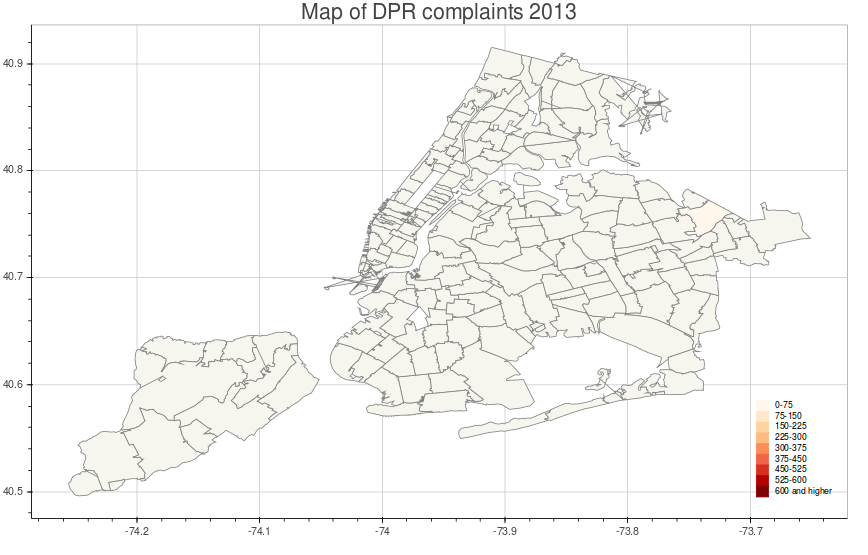


Figure Map of DPR complaints 2013

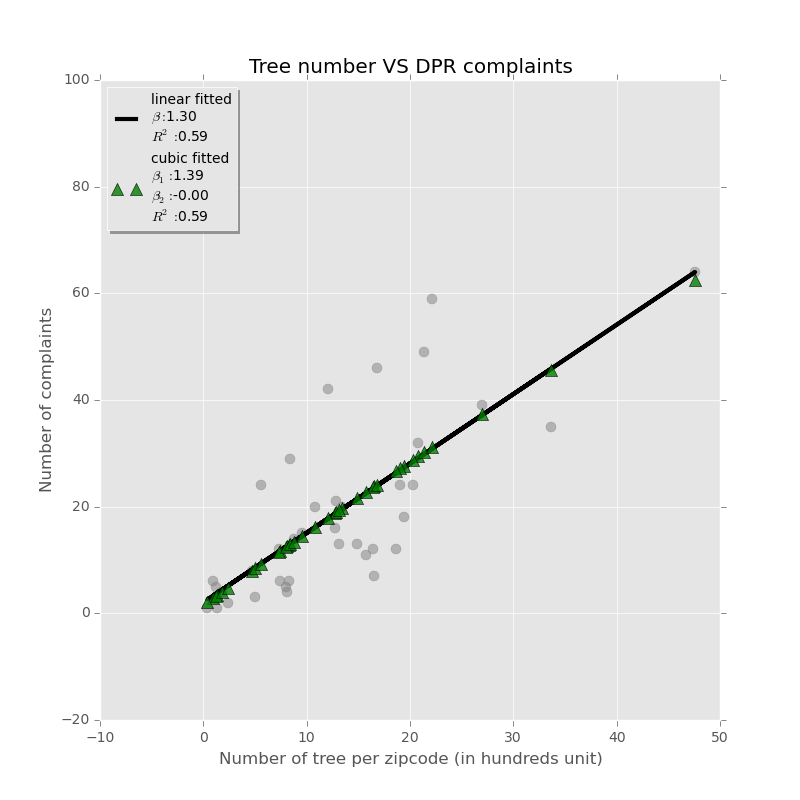


Figure number of trees VS number of DPR complaints 2011

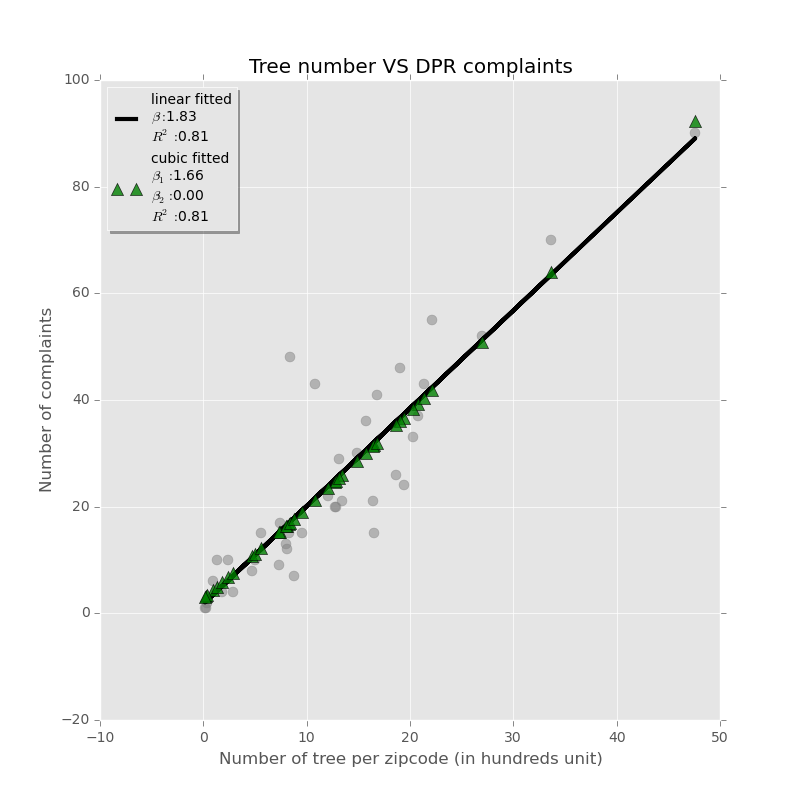


Figure number of trees VS number of DPR complaints 2012

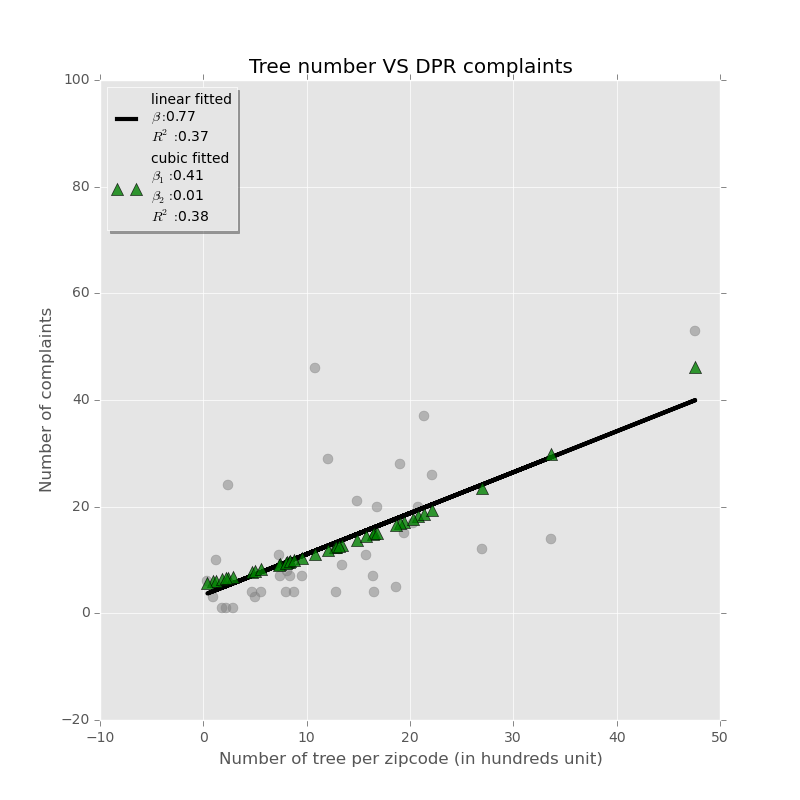


Figure number of trees VS number of DPR complaints 2013

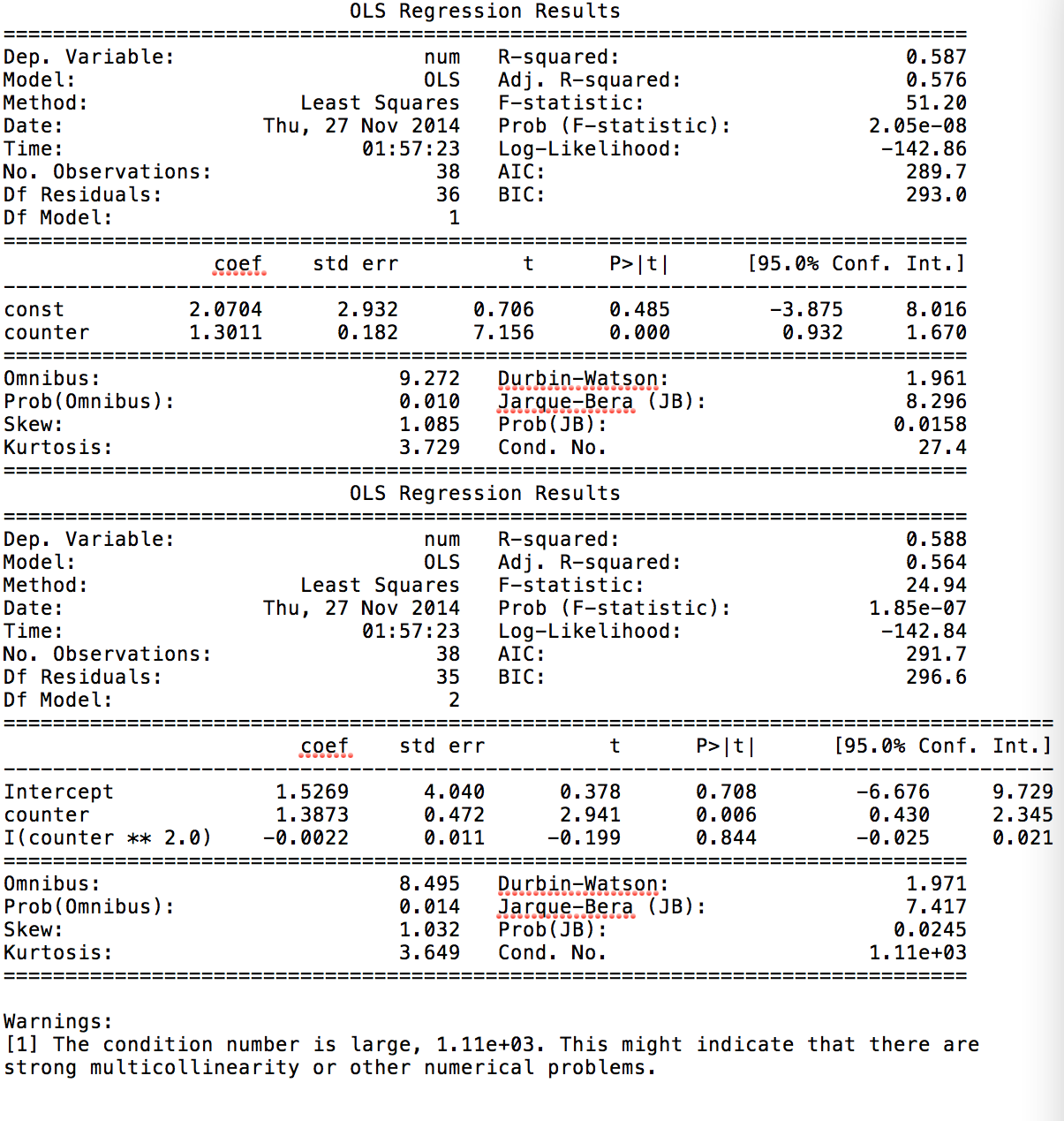


Figure Statistics summary for tree VS DPR complaints 2011

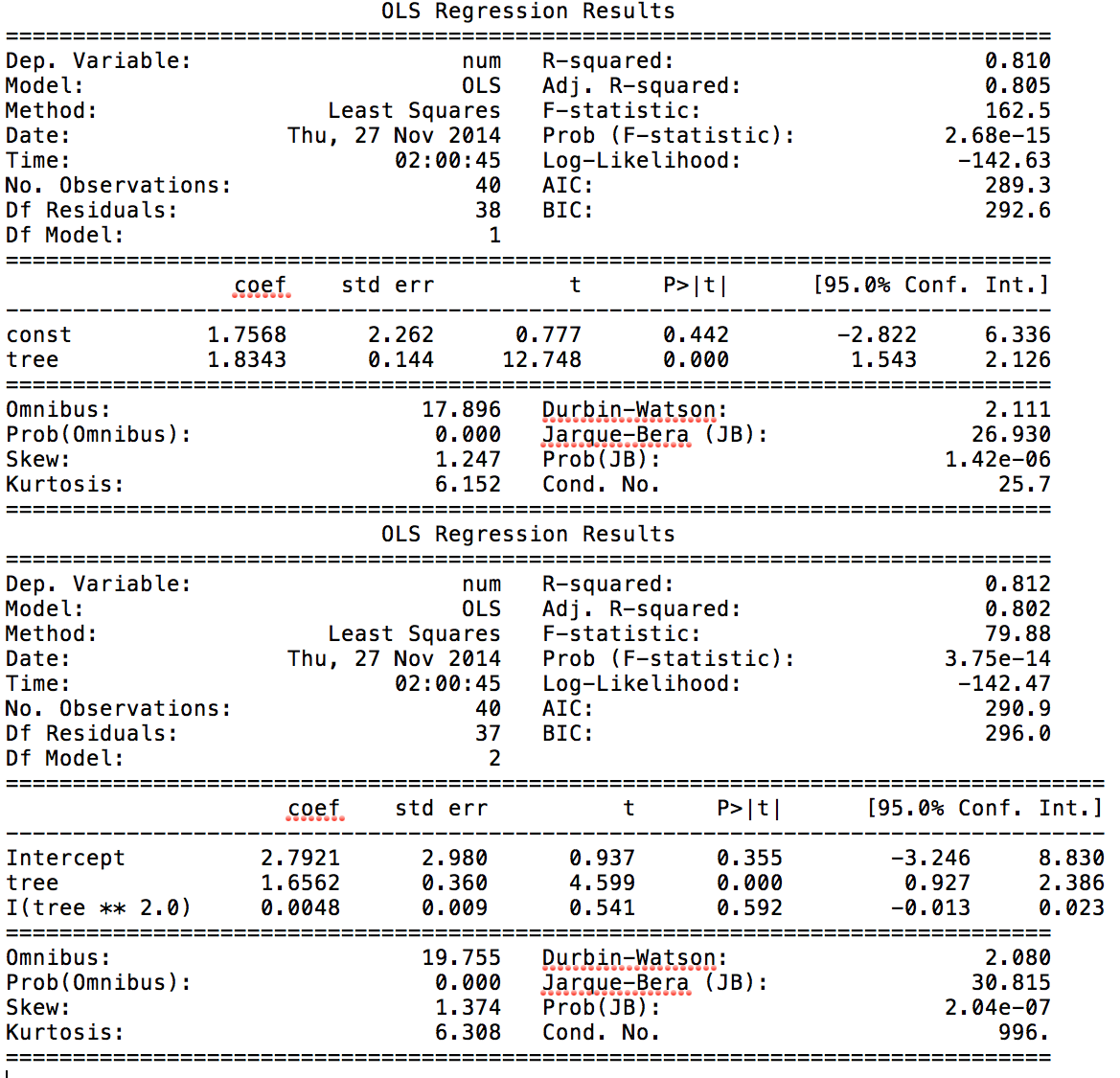


Figure Statistics summary for tree VS DPR complaints 2012

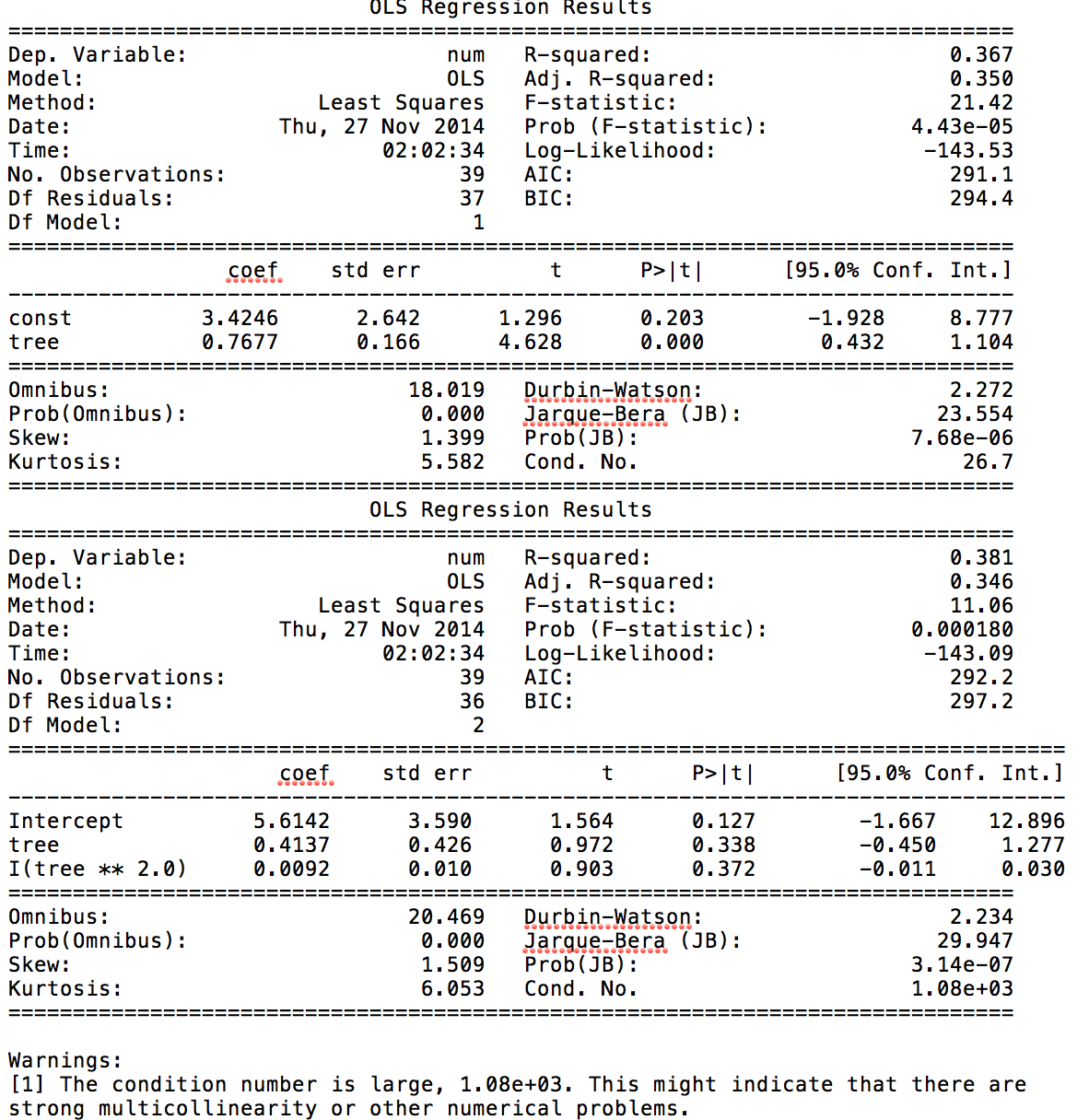


Figure Statistics summary for tree VS DPR complaints 2012

# References

1. Blake, Eric S., Kimberlain, Todd B., J. Berg, Robery, P. Cagnialosi, John, and L. Beven II, John, 2013. *“Tropical Cyclone Report: Hurricane Sandy,”* National Hurricane Center*, National Hurricane Center* February 12, 2013.
2. Bloomberg, Michael, 2013. *A Stronger, More Resislient New York.* NYC Special initiative for Rebuilding and Resiliency, 2013, pp 1-438.
3. Silive 2011, *October snow storm pounds Staten Island with 2 inches*. Last updated on November 03 2011.

<http://www.silive.com/news/index.ssf/2011/10/october_surprise_2_inches_of_s.html>.

1. Luís M. A. Bettencourt and Geoffrey­ B. West, “Bigger Cities Do More with Less”, U.S. Patent And Trademark Office (data on patents filed between 2000–2005 for U.S. metropolitan areas).