University at Buffalo
School of Engineering and Applied Sciences

# Predicting gasoline shortage during disasters using social media

Abhinav Khare, Dr. Qing He, Dr. Rajan Batta
Department of Industrial and Systems Engineering, University at Buffalo
November 6th , 2018


# Outline

- Motivation
- Objectives and Challenges
- Data Description
- Methodology
- Case Study and Results
- Contribution
- Future Work

# Motivation

- Gasoline is an extremely essential commodity pre/during natural disaster
  - Evacuations
  - Generators
- Surge in demand  as people panic-by and hoard supplies → shortages
  - News outlets
  - Word of mouth
- Shortages → people stuck/helpless in high risk zones without essential supplies
- Such shortages became very prominent during the onset and post landfall of Hurricane Sandy (2011) and Hurricane Irma (2017).

## Motivation

- According Florida Department of Transportation, during Irma demand of gasoline went up by 150%.
- There was enough gasoline at the ports to replenish the stations and satisfy the demands.
- There were not  enough drivers and vehicles.  They were brought in from Arizona later.
- People tweeted about these shortages.

"The shelters are full, there is no **gas**. Tornados could happen, and storm surge is predicted. So what are people supposed to do? **#Irma** "

Saw 25 Electrical Boom Trucks & 15 **Gas** Tankers heading south on I-95 today. #greatfeeling think they were headed to Florida **#Irma** #NYtoFLA"

"We are riding it out in Jacksonville Fl many are without power down south (84,000)**gas** shortages so evacuating difficult **#Irma** is no joke 😳🙏"

"My inlaws were going to stay in Port Charlotte until forecast for landfall changed, had barely enough **gas** to get to Gainesville **#Irma**"

"GasBuddy app now supports motorists seeking diesel fuel. **#Irma** #HurricaneIrma **#gas** #Florida"

"Insane..95% of Florida trying to leave at one time. Roads r slammed. No **gas**. No hotels available. Scared to see my neighborhood after **#irma**"

"Gas stations out of gas, water shelves empty, stores and airports closed. Stocked up on food and

wine, waiting on #irma"

# Objective

- Natural questions that arise :
  1. Can social media data, especially twitter be used to predict these shortages?
  2. If so, what methods would solve this problem?

Our objective was to answer these two questions.

- If the demand surges and shortages can be forecasted then the authorities can plan a response to the demand
- Appropriate amount of gasoline can be directed to the shortage affected regions earlier and efficiently.

# Challenges

Challenge 1 : How to identify tweets about shortage ?

- Twitter data is difficult to process and classify.
- It is unstructured, noisy and contains a plethora of information.
- A tweet contains a max of 140 characters, is informal, contain abbreviations and spelling mistakes.
- classifying tweets for a specific problem like identifying gasoline shortage has never been done.
- Identifying important features for this classification task is a novel and unique question

Challenge 2 : How to forecast the actual spatio-temporal shortage from tweets.

- Spatio-temporal distribution of tweets not equivalent to the spatio-temporal shortage distribution.
- Spatial & temporal lag between the origin of the shortage & the tweet about shortage is an uncertain quantity.

# Data Description

- Our data one million tweets from Florida during the period 6-15 September 2017.
- 1048575 rows and 41 columns that include TWEET ID, TWEET TEXT, USER ID , DATE, HASHTAG, LATITUDE, LONGITUDE, BOUNDING BOX

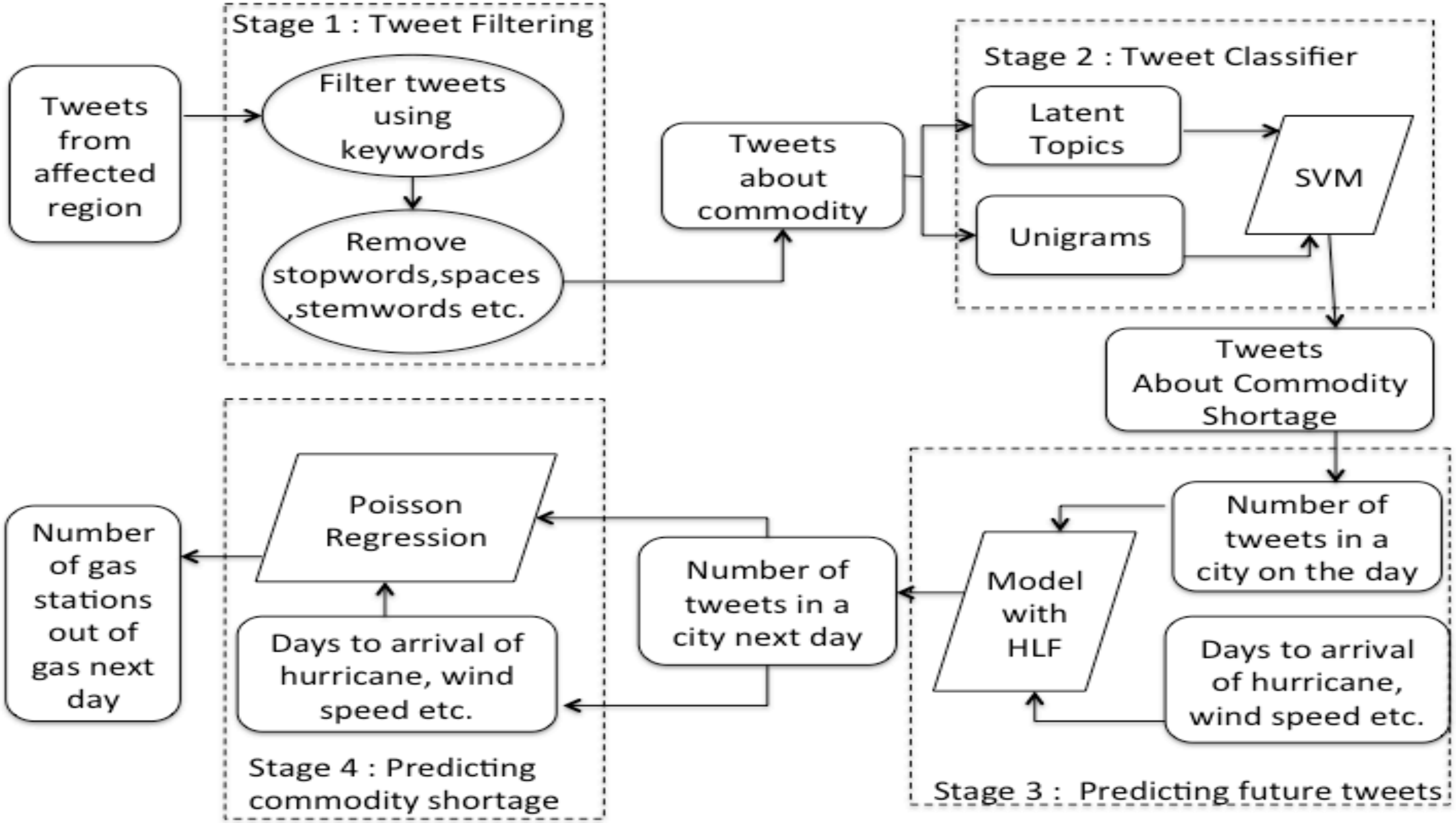
| Summary Statistic | Values |
|---|---|
| Number of Tweets Collected | 1,048,575 |
| Number of Unique Twitter Users | 111,801 |
| Period of Data Collection | 6th Sept 2017- 15th Sept 2017 |
| Date of Irma Landfall in Florida | 9th Sept 2017 |
| Number of tweets prior to Irma landfall in Florida | 456,530 |
| Number of tweets during Irma in Florida | 151,792 |
| Number of tweets post Irma in Florida | 440,253 |
| Number of Gas Related Tweets (labeld manually) | 4070 |

# Data Description

- Collected ground truth about gasoline shortage in Florida from 6th-15th Sept 2017 from Gasbuddy App for validation of methodology
- Collected details and predictions of the hurricane Irma path from from 6th-15th Sept 2017 from the National Hurricane Center website

| City | Date | Proportion of Gas Stations Without Gas | On Hurricane Path | Inside 3-day Cone | Inside 5-day Cone | Days to Arrival | Watch/ Warning | Wind Speeds (mph) |
|---|---|---|---|---|---|---|---|---|
| Gainesville | 09/07/17 | 0.58 | y | n | y | 4 | n | 175 |
| Jacksonville | 09/08/17 | 0.31 | n | y | y | 3 | n | 155 |
| Miami | 09/07/17 | 0.42 | y | y | y | 3 | watch | 175 |
| Orlando | 09/08/17 | 0.35 | y | y | y | 3 | watch | 155 |
| Tallahassee | 09/08/17 | 0.46 | n | n | y | 3 | n | 155 |
| Tampa | 09/06/17 | 0.3 | n | n | y | 5 | n | 185 |
| Naples | 09/07/17 | 0.54 | n | y | y | 3 | watch | 175 |

# Methodology



# Methodology - Stage 1 - Tweet Filtering

1. “Gasoline-related” are the filttered out using from the huge corpus tweets.
   - Used keyword search :any word that contains the “string” gas is a possible keyword
   - Used regular expression with the grep package in R for the keyword search.
   - Hashtag search : ^gas finds words starting with “gas” and also finds words that contain the string “gas”
   - Tweet search : \\bgas finds sentences with “gas” anywhere in the sentence
   - Results from tweet and hashtag search.

2. Tweet cleaning for removing noise and uniformity (for classification)
   - Removed user names, links, punctuations, tabs, general whitespaces, stopwords, and numbers.
   - Changed words to stem words and to lower case

# Methodology – Tweet Filtering – Irma Results

- Key words found using regular expressions:

  *gasoline, gas, gasinmiami, gaspricefixing, gasstation*
  *gasservice, gastateparks, gasshortage, gasoil,*
  *gastation, gaswaste, nogas, outofgas, findgas*

- From 1 million to 4070 gasoline-related tweets were filtered out.

# Methodology - Stage 2 - Tweet Classification

- Classified “gasoline-related” into “gasoline-shortage tweets” & “ non gasoline-shortage tweets”

- Used a SVM classifier with two kinds of features ;
  1. Unigrams or words
  2. Latent Topics identified using 4 topic modeling techniques namely
     i. Correlated Topic Models (CTM),
     ii. Latent Dirichlet Allocation (LDA) using Variational Expectiation-Minimization algorithm (VEM),
     iii. LDA using fixed VEM (VEM Fixed),
     iv. LDA using Gibbs Sampling (Gibbs)

- To find the best set of features , conducted an experiment.
  1. Fixed number of latent topics in the model and varied frequency threshold of unigrams .
  2. Fixed the number of unigrams and varied the number of unigrams in the model.

# Methodology - Tweet Classification – Irma Results

Performance of SVM using topics and unigrams (varied word frequency threshold, number of topics = 5, train/test = 70/30

| Frequency threshold | Number of words | Precision | Recall | F score |
|---|---|---|---|---|
| 5 | 937 | 0.9406566 | 0.7136015 | 0.8115468 |
| 6 | 797 | 0.9604142 | 0.7884615 | 0.8659845 |
| 7 | 710 | 0.9503121 | 0.7620137 | 0.8458096 |
| 10 | 519 | 0.9692899 | 0.7715618 | 0.8591967 |
| 20 | 282 | 0.9636684 | 0.7667436 | 0.8540008 |
| 50 | 109 | 0.9721385 | 0.7757009 | 0.862881 |
| 100 | 38 | 0.9722495 | 0.7793427 | 0.8651735 |
| 350 | 5 | 0.9852071 | 0.7894737 | **0.8765465** |

# Methodology - Tweet Classification – Irma Results

- Performance of SVM using topics and unigrams (varied number of topics, word frequency threshold= 350, train/test = 70/30

| Number of topics | Precision | Recall | F score |
|---|---|---|---|
| 2 | 0.9634378 | 0.7670813 | 0.8541196 |
| 4 | 0.9831247 | 0.7612366 | 0.8580682 |
| 5 | 0.9852071 | 0.7894737 | **0.8765465** |
| 6 | 0.9503142 | 0.7903614 | 0.8629887 |
| 10 | 0.9722495 | 0.7793427 | 0.8651735 |
| 10 | 0.9755556 | 0.71388 | 0.8244524 |
| 15 | 0.9733728 | 0.7878813 | 0.8708593 |

# Methodology - Tweet Classification – Irma Results

- The best F1 score was obtained using 5 unigrams and 5 topics.
- The 5 best unigrams were gas, get, line, out, station.

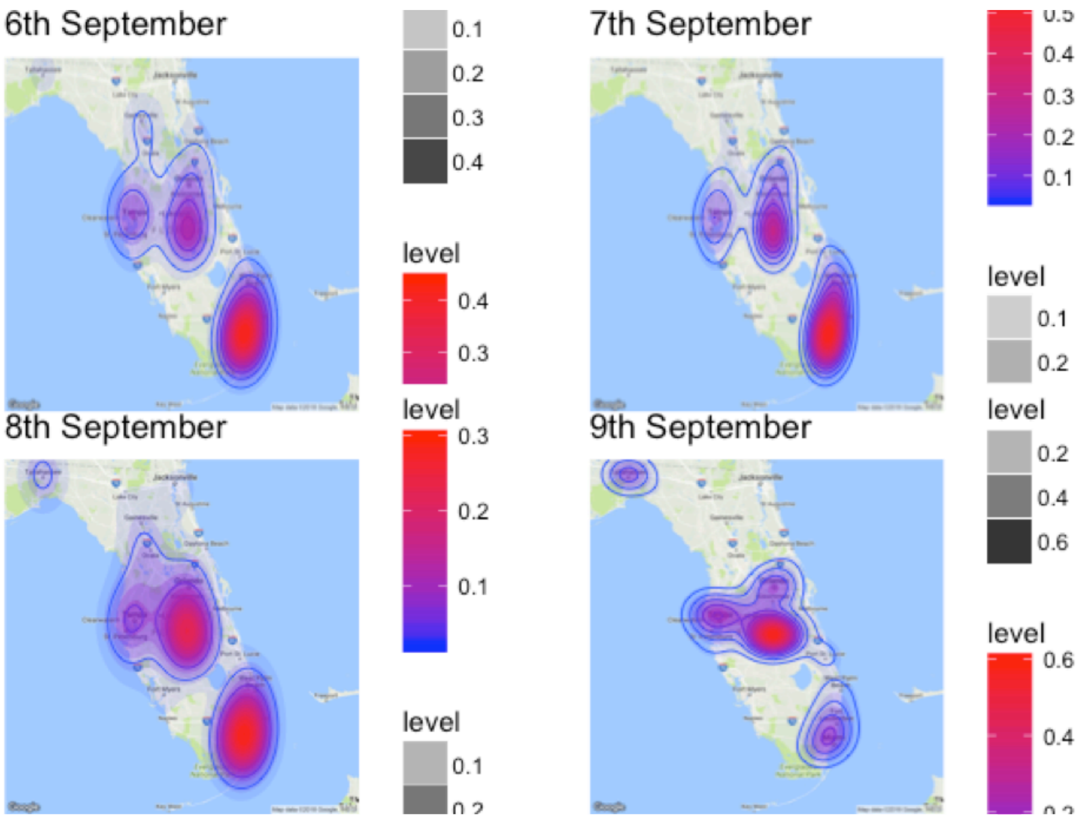
**5 topics identified using CTM topic modeling techniques**

| CTM | | | | |
|---|---|---|---|---|
| Topic 1 | Topic 2 | Topic 3 | Topic 4 | Topic 5 |
| gas | station | gas | gas | gas |
| cannot | gas | no | station | price |
| find | need | station | wait | high |
| know | hurricaneirma | line | line | got |
| where | close | miami | irma | irma |

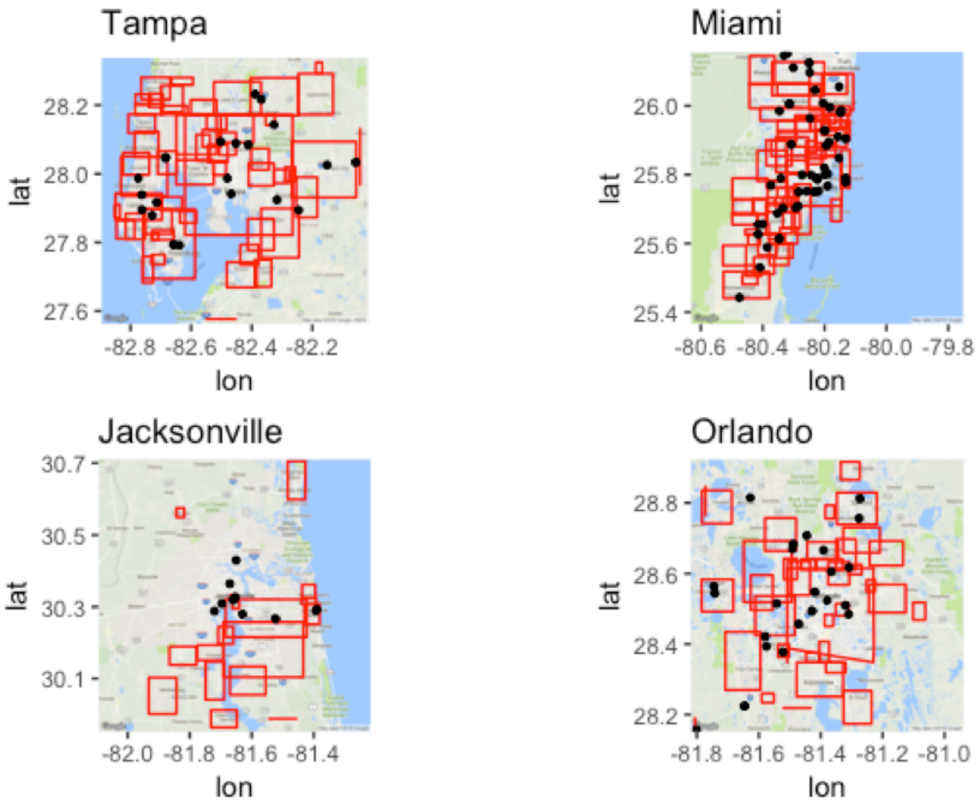

**Word cloud of top 50 unigrams**

# Methodology - Stage 3 – Predicting number of future gasoline-shortage tweets

- To forecast the spatio-temporal distribution of the gasoline-shortage, spatio-tempral analysis




**Heat map of gasoline shortage tweets**



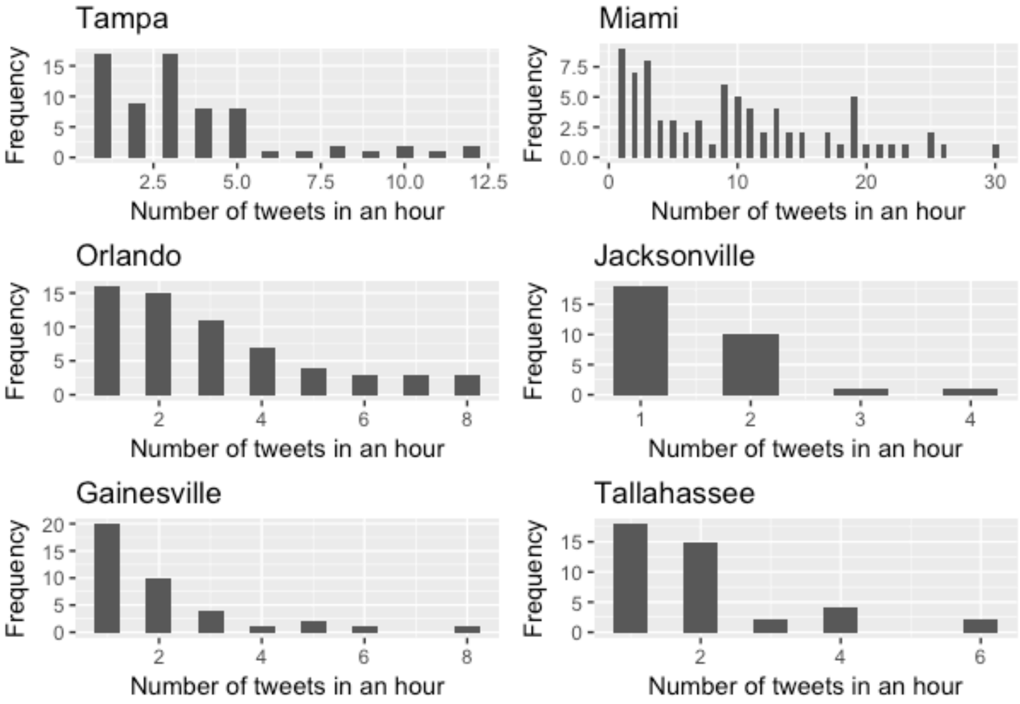

**City-wise geo-location of gasoline shortage tweets**

# Methodology - Stage 3 – Predicting number of future gasoline-shortage tweets

- Arrival of gasoline-shortage tweets followed  a Poisson Distribution

## Results of Chi-square tests for Poisson Distribution

| City | Lambda | Chi-squared | p - value | Chi-squared (sim) | p-value (sim) |
|---|---|---|---|---|---|
| Tampa | 3.637681 | 249.15 | 2.20E–16 | 82.29 | 0.0004998 |
| Miami | 9.558442 | 197200 | 2.20E–16 | 48.273 | 0.004498 |
| Orlando | 3.064516 | 24.202 | 0.00212 | 38.742 | 0.0004998 |
| Tallahasse e | 2 | 21.447 | 0.0006669 | 42.854 | 0.0004998 |
| Jacksonvill e | 1.5 | 16.521 | 0.002394 | 41 | 0.0004998 |
| Gainsville | 2.051282 | 42.072 | 5.04E-07 | 68.282 | 0.0004998 |
| Florida | 25.73171 | 2.26E+10 | 2.20E–16 | 24.366 | 0.9995 |




Histogram for number of tweets in an hour

## Methodology - Stage 3 – Predicting number of future gasoline-shortage tweets

- Arrival of gasoline-shortage tweets followed a Poisson Distribution

- Candidate methods for tweet forecast :
  1. ARIMA/SARIMA models for time series modeling
  2. Poisson Regression Method

- We tried both and to increase the accuracy came up with a Hybrid Loss Function Method which combines the ARIMA and Poisson Regression Methods.

# Methodology - Stage 3 – Predicting number of future gasoline-shortage tweets

- Hybrid Loss Function ( Convex )

$$-L(\theta, Y^{'}) = e^{\theta^T X_{train}} - Y_{train}\theta^T X_{train} + \lambda_1(e^{\theta^T X_{test}} - Y^{'}\theta^T X_{test}) + \lambda_2(Y^{'} - Y_{ts})^2$$

- Gradients

$$-\frac{\partial L(\theta, Y^{'})}{\partial \theta} = (e^{\theta^T . X_{train}} - Y_{train})X_{train} + \lambda_1(e^{\theta^T X_{test}} - Y^{'})X_{train}$$

$$-\frac{\partial L(\theta, Y^{'})}{\partial Y^{'}} = \lambda_1\theta^T X_{train} + 2\lambda_2(Y^{'} - Y_{ts})$$

# Methodology - Stage 3 – Predicting number of future gasoline-shortage tweets

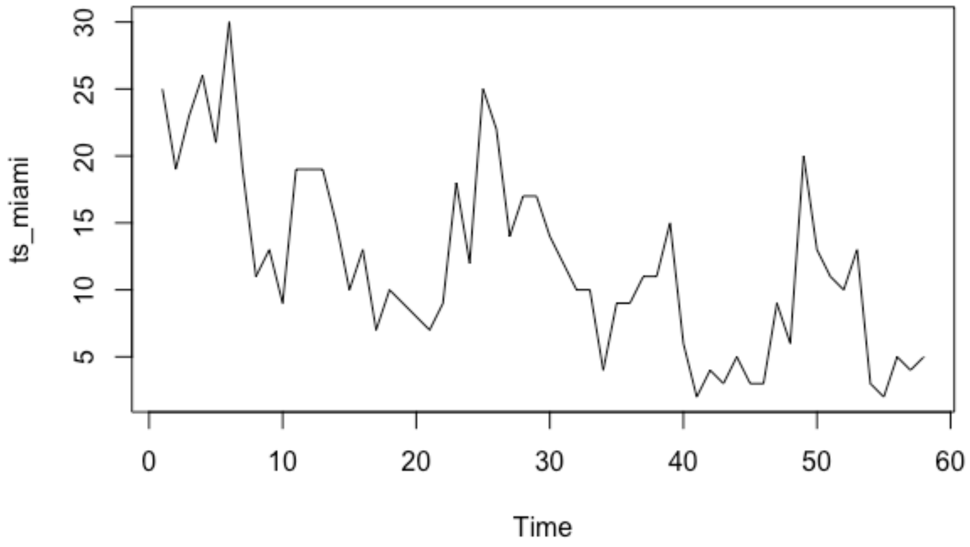
- Poisson Regression fit in predicting future tweets about shortage

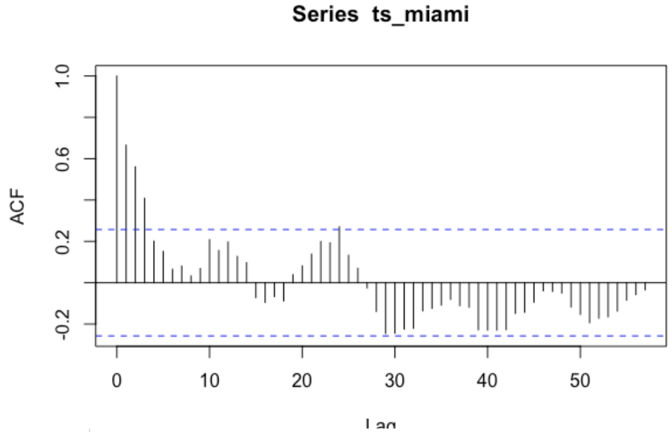
| | Estimate | Std Error | z-value | P-value | |
|---|---|---|---|---|---|
| (Intercept) | -5.632856917 | 1.342732525 | -4.195069988 | 2.73E-05 | *** |
| gas shortage one that day | 10.05553126 | 1.316275539 | 7.639381701 | 2.18E-14 | *** |
| number of gas stations | 0.006732827 | 0.000395052 | 17.04290458 | 3.95E-65 | *** |
| on hurricane path | 0.679578601 | 0.142605366 | 4.765449025 | 1.88E-06 | *** |
| inside 3-day cone | 1.308414331 | 0.143248048 | 9.133906853 | 6.61E-20 | *** |
| days to arrival | -1.71048558 | 0.148740094 | -11.49982855 | 1.32E-30 | *** |
| watches/ warning | -5.763048725 | 0.418491761 | -13.77099686 | 3.81E-43 | *** |
| watches/ warning | -3.698064077 | 0.265035222 | -13.95310424 | 3.01E-44 | *** |
| wind speeds | 0.058446979 | 0.007923946 | 7.375993819 | 1.63E-13 | *** |
| Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 . 0.1 1 | | | | | |
| Null deviance: 2121.05 on 63 degrees of freedom | | | | | |
| Residual deviance: 473.84 on 54 degrees of freedom | | | | | |
| AIC: 707.31 | | | | | |

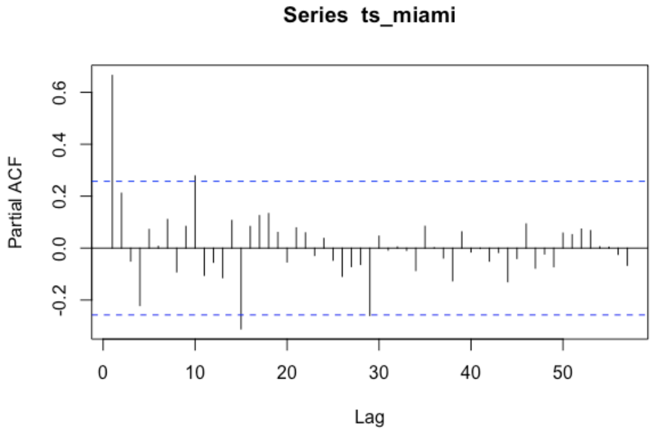
Pseudo-R2 = 0.78

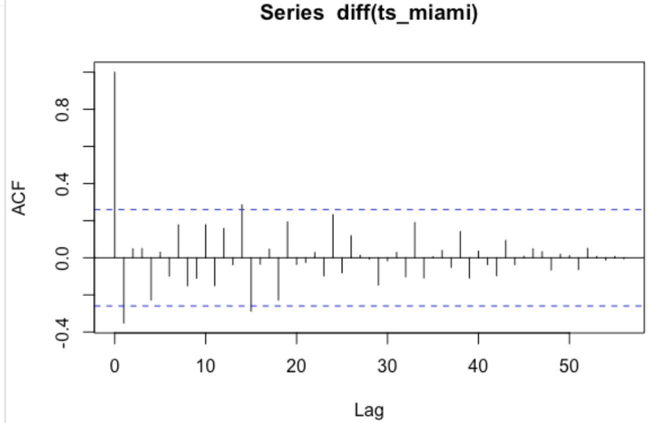
# Methodology - Stage 3 – Predicting number of future gasoline-shortage tweets

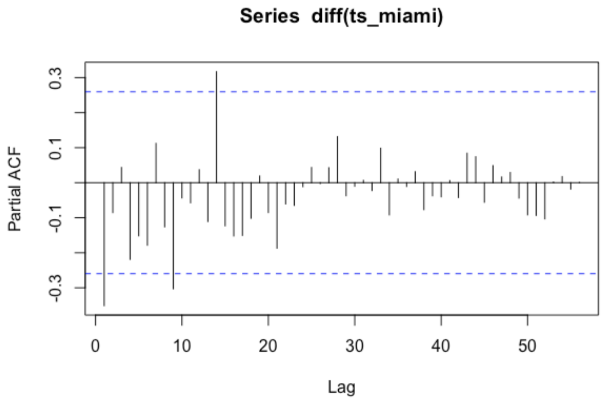
ARIMA modeling for Miami



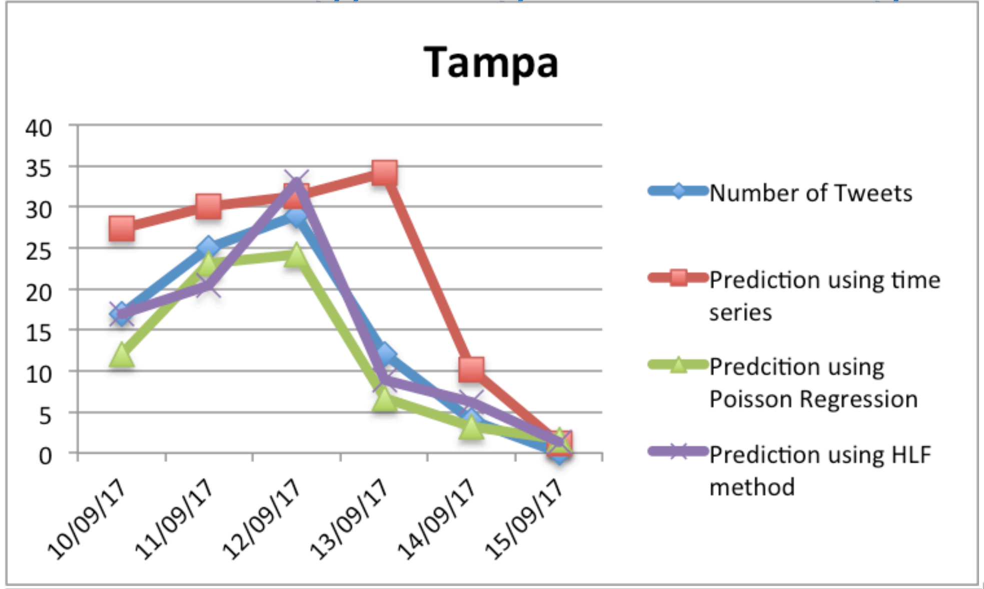



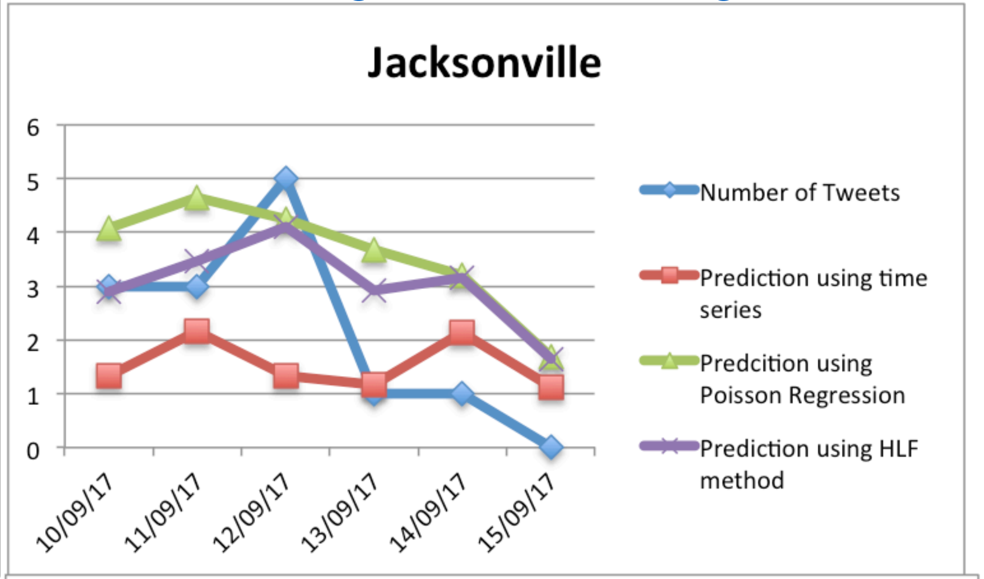



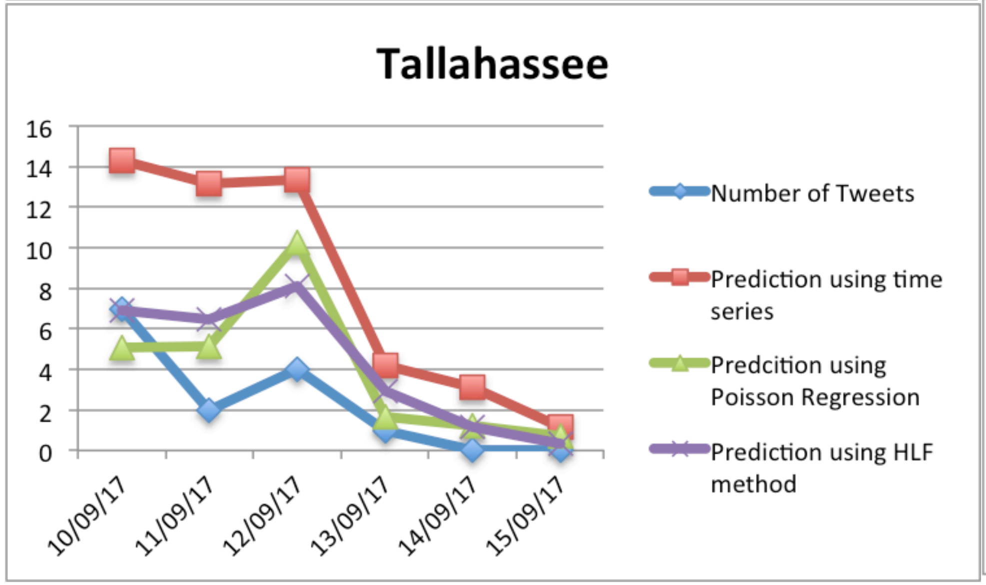



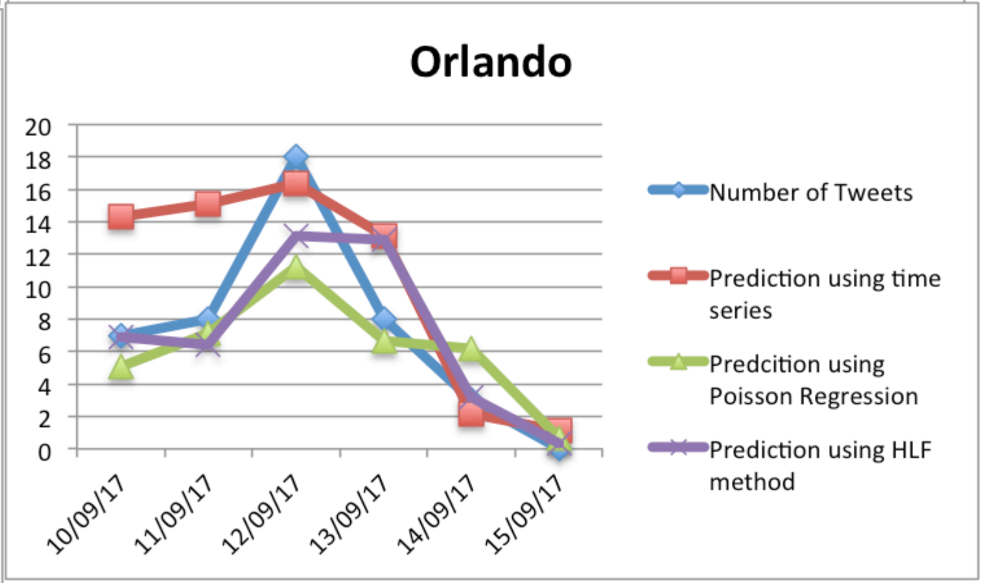

ARIMA(0,1,1) model fitted the data for Miami

# Methodology - Stage 3 – Predicting number of future gasoline-shortage tweets

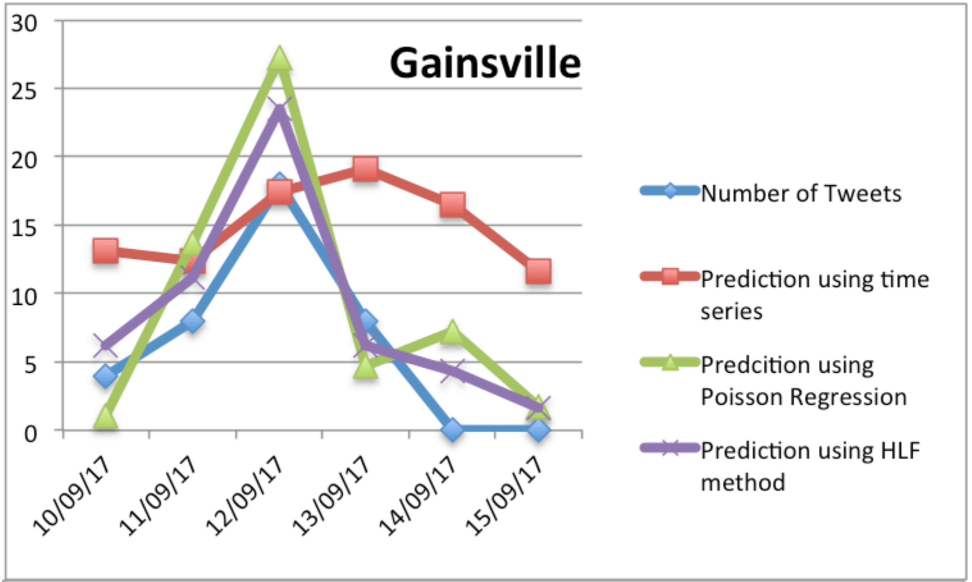


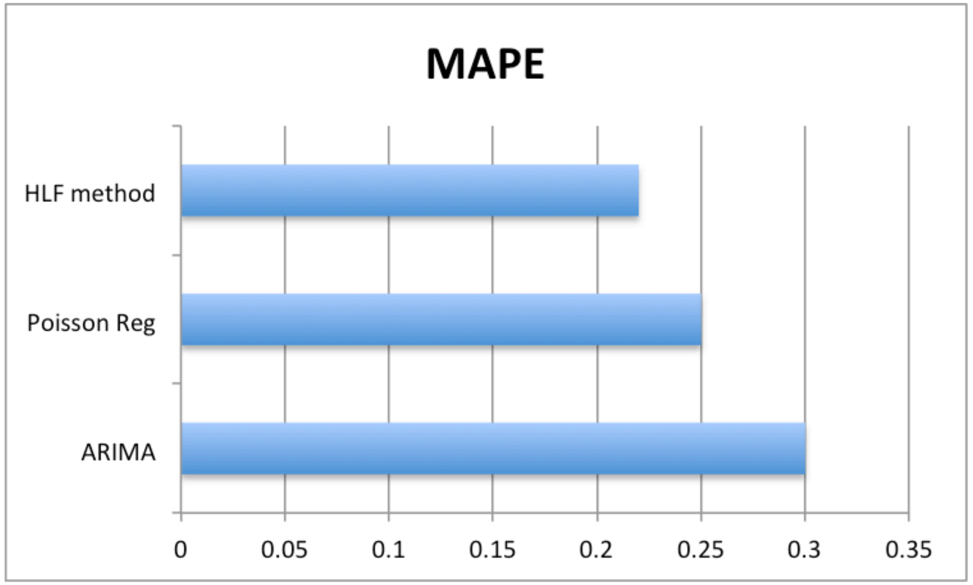



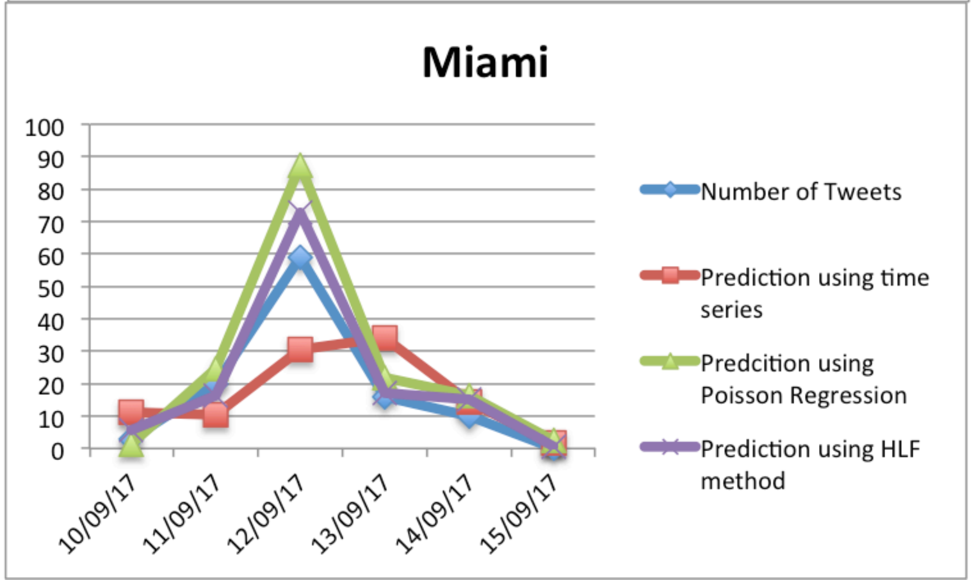



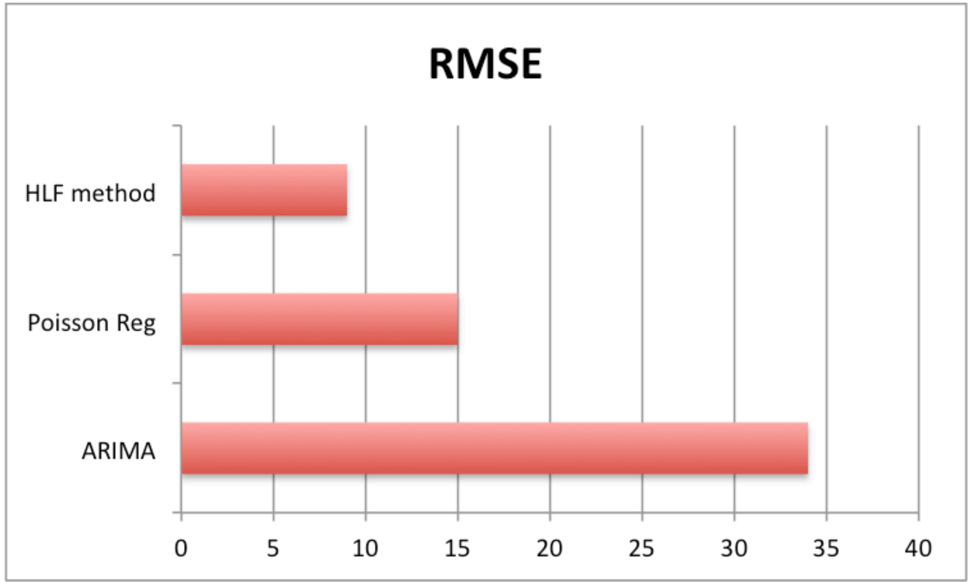



# Methodology - Stage 3 – Predicting number of future gasoline-shortage tweets



# Methodology – Stage 4 – Predicting number of stations without gasoline

- It's a Poisson Regression Model and for Irma data it fit the data extremely well.

| | Estimate | Std. Error | z-value | P-value | Sig |
|---|---|---|---|---|---|
| (Intercept) | 4.255101001 | 0.056026652 | 75.94780113 | 0 | *** |
| Population | -1.18E-06 | 1.35E-07 | -8.700698039 | 3.30E-18 | *** |
| number of gas stations | 0.00310295 | 0.000121362 | 25.56764236 | 3.50E-144 | *** |
| number of tweets the next day | 0.002997528 | 0.000281172 | 10.66083059 | 1.55E-26 | *** |
| days to arrival | -0.137963866 | 0.020294006 | -6.798256761 | 1.06E-11 | *** |
| warningn | -0.20750846 | 0.049436483 | -4.19747623 | 2.70E-05 | *** |
| Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 . 0.1 1 | | | | | |
| Null deviance: 5961.01 on 71 degrees of freedom | | | | | |
| Residual deviance: 689.17 on 60 degrees of freedom | | | | | |

Pseudo-R2 = 0.987

MAPE= 0.31
RMSE= 9.13

# Contribution

- Building of a classifier that identifies gasoline shortage tweets from corpus of all tweets generated.
- Discovering that the arrival of tweets about gasoline shortage follows a Poisson distribution.
- Developing a hybrid loss function method that forecasts the number of tweets about gasoline shortage.
- Development of a four-stage gasoline shortage prediction methodology which takes tweets generated on a day in an affected city as input and generates the number of stations that will be out of gas the next day as the output
- Model validation with a case study based on Hurricane Irma.

# Future Work

- F1 score for tweet classification is good but recall is relatively low compared to precision . Reduction of false positives in the future using a other techniques like recurrent neural nets.
- Our method does a course grain prediction of gasoline-shortage, predicts the number of stations without gas in the city because ground truth about individual stations was not available.
- In the future prediction at individual gas station level if ground truth is available.
- future shortage data is available at the individual gas station level, it can be fed into a decision making model for gasoline delivery to gas stations to ensure adequate supply were it is needed.
- This would likely be a vehicle routing type of formulation.

# Thank you