[[1]](#footnote-2)

Phillip G. Efthimion and Brandon Lawrence

Domestic Crude Oil Inventory Prediction Using Aggregated Twitter Information

*We will use the Twitter hashtag, #OOTT to extract key information relating to the stock price for crude oil, news, floating storage, and tanker movements. The relevant information will then be automatically transferred into our database. Using the scraped data, we will be able to construct a time series and other statistical tools to help us predict crude oil inventory movements.*

*Index Terms*—Enter key words or phrases in alphabetical order, separated by commas. For a list of suggested keywords, send a blank e-mail to [keywords@ieee.org](mailto:keywords@ieee.org) or visit <http://www.ieee.org/organizations/pubs/ani_prod/keywrd98.txt>

# INTRODUCTION

CRude oil is the lifeblood of modern society. It is a unique good in the marketplace as it is heavily traded the world over, can be transported across the globe, stored indefinitely, and ,most importantly, follows the forces of supply and demand predictably. The last point is what this paper is keys in on. Oil revenue is a vital source of income for nearly every developed country, and, with world production at ~80 million barrels per day, the coordination required to meaningfully distort the market is incredible. The Organization of Petroleum Exporting Countries (OPEC) managed to control nearly 40% of the world petroleum supply for a few decades, but their spectacular falling out in 2014 due to pressures from U.S. shale producers really exposed how tenuous their power really was.

With the fall of OPEC came more reliable market responses to supply and demand shifts, making inventory data much more valuable. Surprisingly, the methods of tracking oil inventory movements are quite manual. Some traders rent spaces overlooking key ports and observe how low ships are sitting in the water to gauge movement levels.

In 2015, a group of users on Twitter began sharing their information using the #OOTT tag. #OOTT stands for the Organization of Oil Trading Tweeters. It acts as an aggregator for market research and news regarding the oil industry.

Using MySql, XML, the Twitter API and the Python package Tweepy, we will take data aggregated on the hashtag and use it to predict future domestic crude oil inventory movements with time series analysis. We will also look to see what variables are significant to the price of crude oil with variable selection methods.

The U.S. Energy Information Administration (EIA) is a government agency that provides data on a wide range of energy industries, such as gas, solar, and wind. They are one of the sources that provide information weekly that relates to the price of crude oil, which is then uploaded to #OOTT on twitter. This will be one of our primary sources of information that we wish to scrape from the hashtag as it comes from an official source.

# Data Gathering

There are many hashtag tracking tools already for download on the internet, both free and by subscription such as Keyhole and hashtracking.com. However, research into these tools show that these provide only analytics such as the amount of activity such as the amount of retweets, favorites, or views a tweet gets. Also, none of them are specified to a specific hashtag and are more about comparing hashtags to see the most optimal way to spread one’s message. In contrast, what we are more interested in is what content is being placed in the hashtag. Though we will eventually use the view data of the tweets to compile user influence ratings for further research, we are more interested in the contents of the tweets. We want to know that within the specified hashtag #OOTT, what users are talking about and the most common topics of the day. This will gives us a good aggregate for what people in the oil industry are discussing and what data being posted there is arguably the most valuable and should be read.

We have used Tweepy, a Python wrapper for the Twitter API interface, to scrape live tweets and to search historical tweets. We looked at historical tweets in order to build our model, already knowing the crude inventory movements for that time period. As we build a more robust statistical model using machine learning, we will begin analysis based on live tweets.

All of the response data was gathered from the U.S. Energy Information Administration (EIA). The EIA is an independent agency that is part of the federal statistical system. It has long been considered a quality place to get unbiased information. The EIA publishes total stocks for petroleum and other liquids weekly with a week lag. Crude inventories are influenced by numerous factors, so we needed to determine the biggest leading indicators in order to focus our model and tighten our analysis. They also have historical data available in an excel file for download going back several decades. We access the EIA data via the agency’s RESTful application programming interface. The statistical analysis for this regression was performed using R, a statistical program.

## Abbreviations and Acronyms

OOTT is the Organization of Oil Trading Tweeters. A popular hashtag on [twitter.com](http://twitter.com) that aggregates oil pricing data. SPR is the Strategic Petroleum Reserve, an emergency supply of petroleum housed by the Department of Energy. RBOB stand for Reformulated Gasoline Blendstock for Oxygen Blending., which is simply unleaded gas futures. CBOB is Conventional Gas Blending Components. GTAB is Gasoline Treated as Blendstock, non-certified Foreign Refinery gasoline classified by an importer as blend stock. LRG is an abbreviation for Liquefied Refinery Gases. NGPLs is an abbreviation for Natural Gas Plant Liquids. They are hydrocarbons in natural gas that are separated as liquids. OPRG is Oxygenated Fuels Program Reformulated Gasoline. PAD is the Petroleum Administration for Defense. These are 5 districts set up for allocating oil. RFG is Reformulated Gasoline..

## Other Recommendations

To begin our analysis, we will look at explanatory variables provided to us by the EIA’s weekly reports. It lists commercial crude oil excluding lease stock, commercial crude oil including lease stock, Alaska in-transit, and SPR as factors that affect the price of crude oil including SPR. However, we are going to look at other factors as well such as total crude oil and petroleum products both including and excluding SPR, total motor gasoline, fuel ethanol, kerosene-type jet fuel, distillate fuel oil, residual fuel oil, and propane/propylene as factors that could affect the price of crude oil. Other explanatory variables that we will look at are stock prices from Fortune 500 oil companies as well as the GDP of countries that export large amounts of crude oil to see if there is any statistical significance to the price of crude oil. Again, all of this information will be provided from #OOTT.

The ebb and flow of crude inventories send strong price signals to the market. Getting an idea of the sentiment before official release can result in a good idea of where crude prices will be heading and possibly provide greater insight into the wisdom of the aggregated global oil trading community.

# Analysis Overview

In order to better understand the weekly crude oil data from the EIA, we determine which of the almost 40 variables we would consider to be statistically significant in order to create a model. This model will hopefully help us predict the future stock level of crude oil. This information will be useful for buying and selling futures contracts and energy stocks.

Multiple linear regression is performed in order to determine which variables are statistically significant. The first issue we run into is that there are not values for all of the variables throughout all of our data. Some values were not tracked until recently while others stopped being tracked at all. There needs to be a value in each column for each row of data to be used in regression. Therefore, some variables had to be dropped. As we want to add to this database as more prices are published weekly, variables that no longer were tracked were dropped. Also, in order to include the maximum amount of variables that information was still published about today, we had the regression start much later in the time series. In total, the regression was done with still over 30 of the variables over just under a decade’s worth of data.

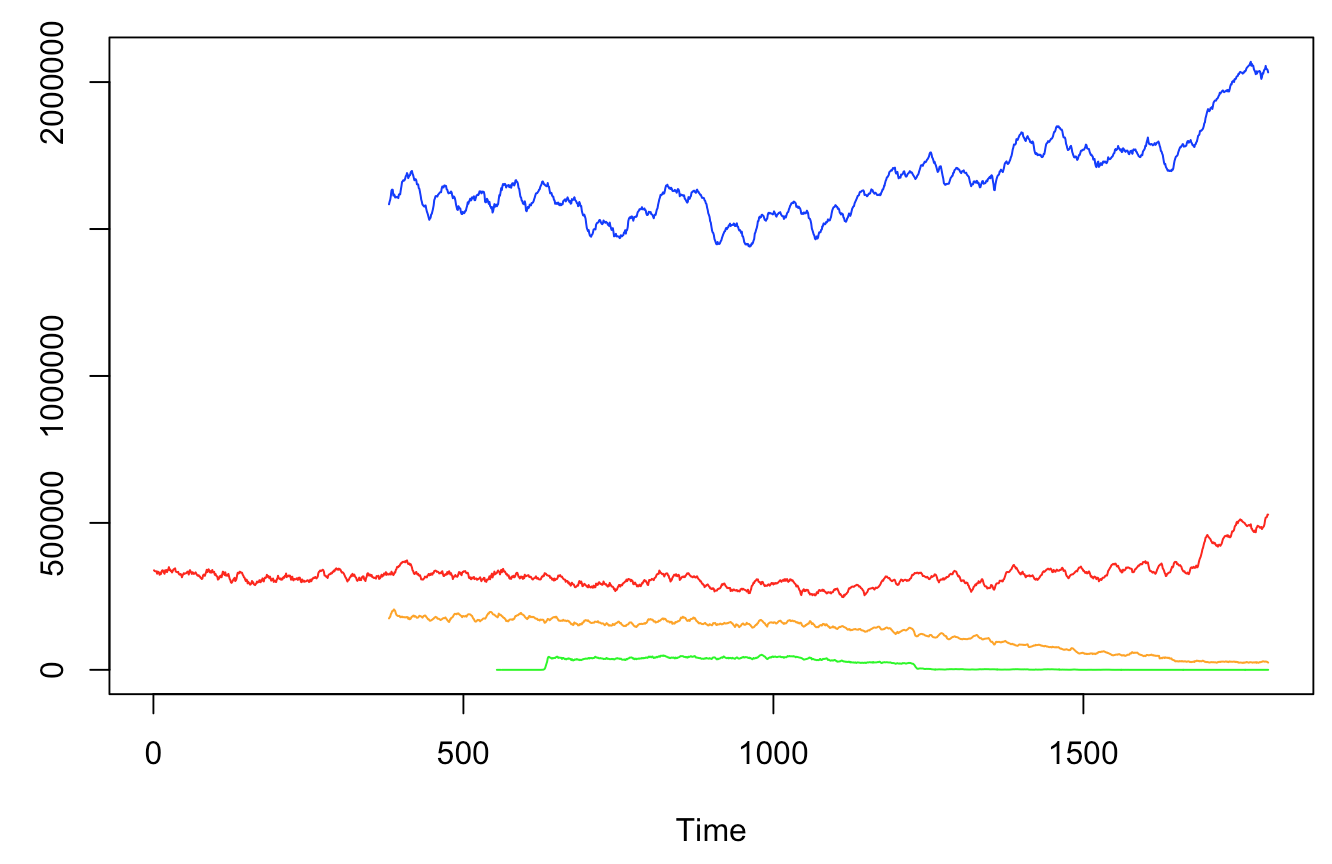
Using an alpha of 0.5, our statistical software tells us that there are 6 variables that are considered statistically significant. However, as we check for covariance, we see that there are really only 4 of our original 40 variables that are considered statistically significant to the price of crude oil. These are the weekly ending stop prices for crude oil and petroleum, crude oil excluding SPR, finished motor gasoline, and reformulated motor gasoline.

This would give us the following equation:

The SPR referenced here refers to the strategic petroleum reserve according to [eia.gov](http://eia.gov). Their definition for the strategic petroleum reserve is “petroleum stocks maintained by the Federal government for user during periods of major supply interruption". This equation tells us that all of these variables have positive effects on the weekly ending stock price of crude oil. We know this because all of the estimates for each variable are all positive. We also can see that the price of reformulated motor gasoline has that largest effect on the price of crude oil.

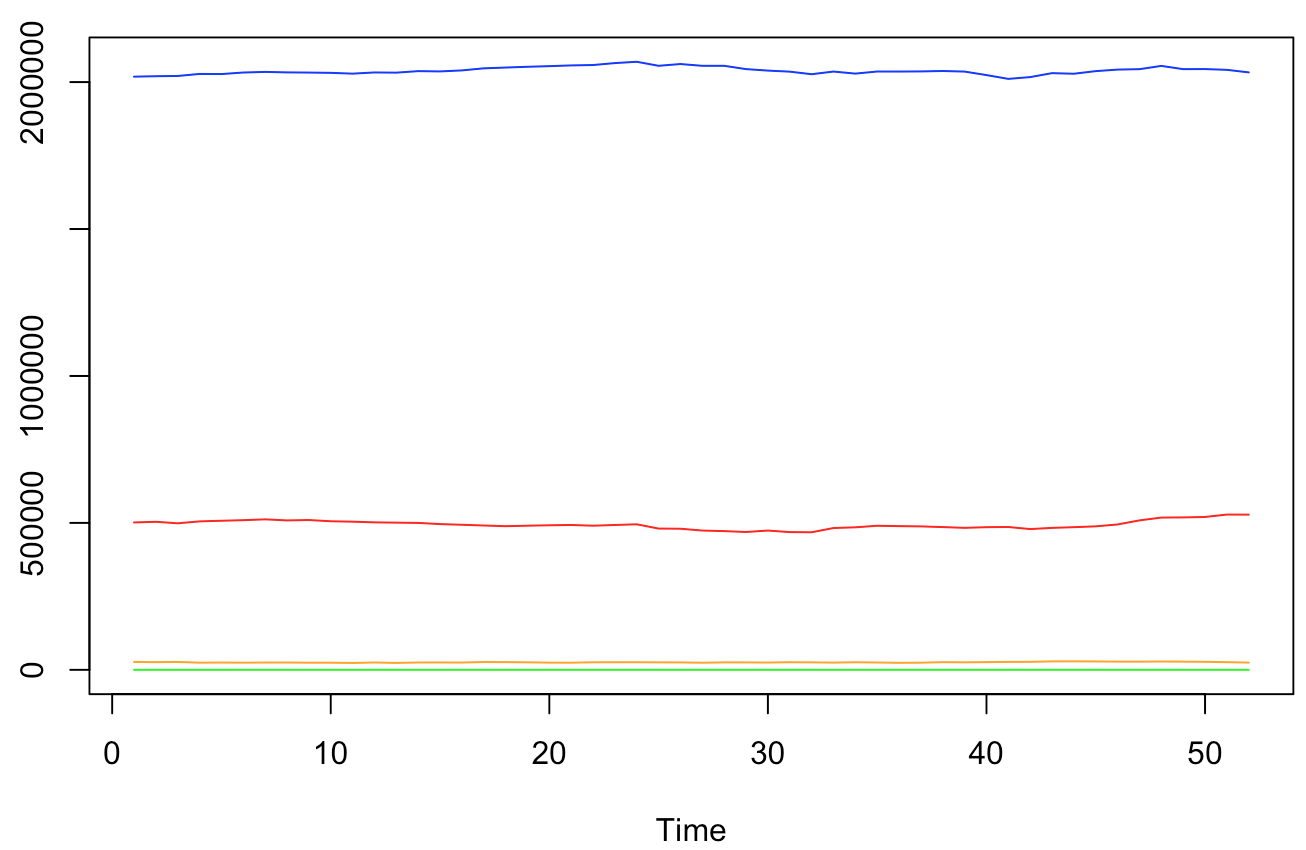
Time series graphs are part of the output of our program. There are graphics that display the historical time series for all of the statistically significant variables of our response variable crude oil as well as a time series for just the past year of data.

Our program uses our regression equation above to show us the time series of each of the statistically significant variables. These time series graphs and the regression were all created using R.



The above graph shows the time series for the variables in our time series: Crude (blue), Crude excluding SPR (red), gasoline (green), and reformulated gasoline (orange). This time series shows us over time how the values increase and decrease over time.

We also output a graph to show just the last year, 52 weeks, of data.



This graph has the same variables as the above graph, but the scale is different. The first graph showed the lifetime of the EIA data. This graph only shows the last year. Another graph available is the data being plotted with a frequency of 52. This provides us with this week’s prices for each year of data. For example, it could provide data for the first week April every year. This helps us see how our data is performing compared to it was one, five, ten, or even more years ago.

Besides outputting graphs, R also provides us with the numbers themselves. These are used for a more quantitative analysis.

Each week, the new week’s data from the EIA website will be scraped and inserted into our database. From there, this new data will be used to provide us with the week’s data that can be used to make financial decisions and add to our time series model.

In future development, there are three expansions that would like to add. The first is comparing this data to how the stock prices performed on a given week historically. Also, the ability to compare how the stock price of crude oil is performing in the United States compared to other regions in the world. The EIA also has this data available in other files for download and would make for a great comparison. Another addition to the database that we would like to make it training our model. Using machine learning techniques such as cluster analysis, we seek to be able to train our model to be more accurate. This would help increase the validity of the model. The final addition that we would make to our analysis of the data from the EIA is the ability to toggle which variable one would want analysis of. One would like to be able to want an analysis of a variable other than crude oil in the data set, such as “fuel ethanol” for example, and automatically the program would perform the statistical analysis to determine which variables are statistically significant and then create time series for those statistically significant variables.

One other large change that we would consider making is the type of database that we are storing our data on. Currently, we are using a MySQL database. However, because of the immense load of data that is inputted into the database from Twitter, NoSQL databases must be considered. If we were to switch to a NoSQL database, we would require a document storage database such as MongoDB. It is scalable and easily fits our needs. Also, MongoDB is able to handle big data like Twitter would provide it with.

Time series analysis will allow us to visualize how the stock price for crude oil has changed and been affected by factors such as seasonality. Time series forecasting will be used to predict wheat future stock prices will be. We will also look at what explanatory variables have the strongest effect on our response variable, the inventory change of crude oil. The AR(1) model will also be used to look for lag and autocorrelation.

Multiple linear regression model will also be created in order to perform variable selection. Using variable selection, we will be able to have an idea of which of our variables are statistically significant to the price of crude oil. We will be using the Durbin-Watson test and the variance inflation factor (VIF) to access this.

Tweepy is a Python package that lets one easily scrape information from Twitter via Twitter's API. We will specify in the API to only take information from posts containing the hastag #OOTT. This will filter the website into what we believe is a suitable aggregate of information that we can process information from. We will be using this and XML to gather our data. Python will be used to scrape the data from Twitter. The data we use will be stored in a database created in MySQL.

The MySQL database consists of the following tables:

Raw\_Tweets - storage for raw API output for user, tweet datetime, and tweet content for a given API query

Tweet\_Repo - clean repository of all #OOTT tweets, the primary table for data analysis

User\_Strength - table that keeps track of individual user weights determined by tweet frequency, predictive power, and interactivity (likes, retweets, replies, followers)

EIA\_Data - EIA data tracking of prices for all crude grades weekly

To store the data from twitter, a database was created using MySQL. There are two tables that our data from twitter is inputted to. The first is the table “tweetstorm”. This table has four columns: id, tweet\_time, username, and tweet. The id is the primary key used for uniquely identifying each tweet, tweet\_time is a datetime type that tells when the tweet was tweeted, the username holds the username of the account that tweeted to the hashtag, and the tweet column contains the 140 contents of the tweet.

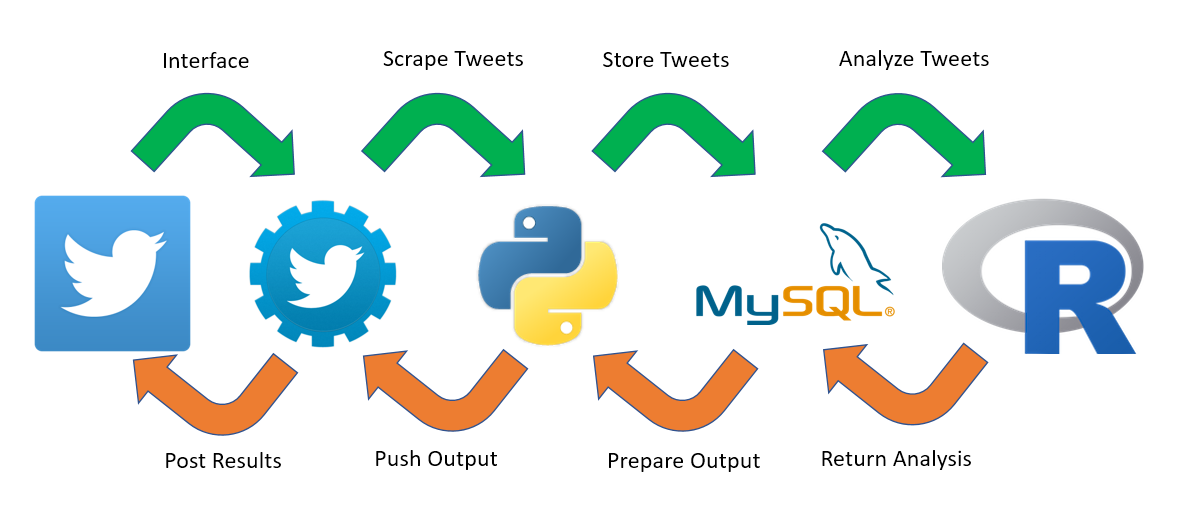
One extra thing that had to be done with this table, and our database systems is changing the variable set. MySQL’s default character set is inefficient for social media information because of the common use of emojis. MySQL’s default character setting does not account for emojis, which are prevalent in social media sites such as Twitter, hence the change to utf8mb4. This table allows us to filter tweets made at specific times, see who are “power users” and tweet most often, and examine the contents of tweets to examine trends from key words.

For example, there was interest for data on activity in the hashtag on Mondays. Monday was chosen as many people catch up on news that they may have missed over the weekend at work or school on a Monday. First we determined what hour was there the most tweets, then within that hour we looked at users who were contributing the most, and what were the most popular topics of the tweets at this busy hour.

To further illustrate this example we tested our queries for data on Monday, March 20th 2017. This returns 24 rows (one for each hour) and the output shows that the hour with the most tweets using the prescribed hashtag is 7:00 AM Central Time. This makes sense, as this would be the 8:00 AM hour on the east coast and New York City, which is a financial capital. From there we see who is tweeting the most at that hour to find the most avid user. From there, we list all of the tweets from that user over the hour. In MySQL, this can be done with the HOUR command to take a specific hour from our ‘datetime’ variable. Running this code daily will let us build a larger picture into when data is available and let us develop graphics of trends showing who power users are and the best time to look for specific information. Plus, what posters should be trusted more with relevant articles they are sharing. This will yield us more information as we collect more data from Twitter.

The second table information was stored in was the table “twitterdata”. This table had some columns in common with our first table such as columns containing unique primary keys, when the tweet was posted, and tweeter’s usernames. However, where this table differs is instead of storing each tweet, it provides binary responses of whether a keyword was used in the tweet, one being ‘yes’ and zero being ‘no’. An example is that if a tweet containing hashtag OOTT contains the phrase ‘Saudi Arabia’, under that attribute for the tweet there would be a one to signify that the tweet contains the words ‘Saudi Arabia’.

These two tables can be joined together in MySQL to show what the tweets that got the most exposure, in terms of retweets and favorites, were. Another example of the table’s uses is to find which of the attributes are mentioned together. For example, we can answer questions such as is ‘futures’ being associated more with ‘higher’ or ‘lower’ today? This is information that could prove to be very useful for people in oil trading.



Sample Workflow

# Concerns & Possible Applications

One potential issue that we would have to look out for in the future is changes that Twitter is making. The company has recently transitioned their replay structure to not count the usernames that one if replying to against the 140 character limit. This has caused some disarray in the discourse on   
Twitter, with spam easily co-opting popular threads and hashtags. We will need to develop a way to more efficiently screen these false tweets out since it is very possible that such nonsense could slow our algorithms down considerably. If this were to be implemented, parameters of the database would need to be altered.

One source of Twitter’s revenue is access to the data and content posted on its site. Twitter only allows one to download 5 days worth of data. In order to get full historical data on the activities and traffic of #OOTT, we would need to compensate Twitter monetarily. So for now, we are downloading the tweeter data of the activity of #OOTT every 5 days to start our own registry of data for our own use. We reasonably expect that most of the tweets will have all of the favorites and retweets that they will get by that point. Purchase of historical data may be considered in the future. With the historical data we would be able to be more accurate about how often certain subjects are talked about and create a time series to visualize this data in the future.

1. P.G. Efthimion, is with Southern Methodist University, Dallas, TX USA. (e-mail: pefthimion@smu.edu).

   B.R. Lawrence, is with Southern Methodist University, Dallas, TX USA. (e-mail: lawrence@smu.edu). [↑](#footnote-ref-2)