Time series analysis and forecast of Benzene levels in gasoline

Introduction

Gasoline releases from the underground storage tanks are frequent causes of groundwater contamination. The nature and magnitude of contamination from such releases depend on the composition of the gasoline. The gasoline composition in US varies spatially as well as temporally for various reasons which include crude oil sources, refinery equipment and capabilities, fuel performance characteristics, industry standards and regulation.

The Clean Air Act Amendments (CAAA) of 1990 (42 U.S. Code 4701) expanded regulation of fuels initially set by the Clean Air Act 1970.The CAAA introduced several requirements that have had a major impact on gasoline composition throughout the United States, beginning with implementation in 1992 and 1995, and continuing to the present. The most important requirements were the total ban on lead in gasoline, and new requirements for three types of gasoline: conventional (CG), reformulated (RFG), and oxygenated (OG). Both reformulated gasoline (RFG) and oxygenated gasoline (OG) required oxygen-containing additives to facilitate clean burning of fuel. Gasoline manufacturers and refineries have complied to these requirements by using lead, different aromatics, alcohols and/or ethers in the gasoline at different periods of time. Limits of benzene, ethers and alcohols concentrations, and vapor pressure are set by these regulations. These components are primarily regulated for their impact on air pollution. However, components like benzene and other aromatics, methyl *tert*-butyl ether (MTBE) and other oxygenates are known ground water contaminants and more importantly, injurious to human health.

The gasoline release events are usually not discovered immediately. When discovered, it is often very expensive to measure the level of contaminants. They are derived by physics based modeling (advection-dispersion models) of the groundwater and the contaminants. Such modeling is highly dependent on the original concentration of constituents in gasoline. Since, the concentrations of these contaminants vary spatially and temporally, it would beneficial to analyze and build a time series model of the contaminant concentration. This would allow to get a precise estimate of the concentration in the past as well as to forecast it into the future. Such models will not only aid in the process based modeling, it will also help determine the effectiveness of the regulations in achieving their objectives. Thus main objectives of this project are:

1. To analyze the contaminant concentration in gasoline to examine the effectiveness of the regulations.
2. To build a time series model to estimate and forecast contaminant concentration

Methodology

The concentration of a specific contaminant, Benzene was analyzed in a specific regulation zone that consists of New York, New Jersey and Connecticut. The RFG program was implemented in this zone and it limited the amount of benzene and total aromatics in reformulated gasoline. The limits were met either on per-gallon basis(<1.00 % vol) or averaged basis (<1.3 % vol). These

limits were enforced starting 1995. The Benzene content in all US gasoline was limted to 0.62 % vol by in 2011 by the Mobile Sources Air Toxics Rule.

A long-term data between 1976 to 2017 was compiled from three different sources and was available as csv files. The dataset contained concentrations of different gasoline components and measurements of different gasoline properties acquired at different points of sale throughout the regulation zone. The analysis in this project was limited to Benzene concentrations measured from 1995 to 2017.

Three different time series forecasting methods were used to develop time series models, Simple Exponential Smoothing (SES), Holts linear trend method and Seasonal Auto Regressive Integrated Moving Average (SARIMAX) model. The data between 1995 and 2010 was used to train the models and data after 2011 was held back to test the models. All the modeling was done with python and it's relevant packages. The Jupyter notebook was used as the coding and presenting platform.

Data wrangling

The data was collected from three different sources. The data was first compiled into a csv using a Java routine. This routine was not included in this project. The csv file was saved as a excel spreadsheet. The spreadsheet contained 53 columns and 22657 entries. The spreadsheet was converted into pandas Dataframe called 'raw_data'. Then the columns/variables which were supposed to be numeric type were identified and list 'colKeys' was created with these column headers. A function 'cfun' was defined to strip the string entries off their first element and return rest of the string. This was done to remove any '<' symbol before the entries. Then the string entries were converted to numeric data types. The entries with 0 values were basically non-dictated data and hence replaced with a NaN.

The dataframe 'raw_data' contained some columns with no data. These columns were removed from the dataframe. The index of the dataframe was set to 'datetime' format. Then the columns with numerical datatypes were identified and listed in num_col. The time series of these columns were plotted to visualize the data. The benzene data was isolated for further analysis.

Data Story

There were total 33 columns with numeric data types. Each column contained time series data for different constituents and properties of gasoline. The time series spanned from1976 to 2017. However, availability of data was not uniform across all the constituents or properties. For example, data on Manganese, ETBE, DIPE, T-Butanol, Propanols, Pentanols, Methanol were sparse and probably not very useful. Similarly, data on MTBE, TAME, Methanol, Ethanol was available for limited periods. There also seemed to be duplication of data as some constituents concentrations were available as percentage of volume and percentage of weight.

The time series plots revealed changes in regulatory requirements. For example, Benzene concentrations decreased quite noticeably after 1995. Lead concentrations decreased gradually and disappeared altogether after 1985 which was when a complete ban on lead was enforced.

The ethanol concentrations rose after 2000 and became steady after which reflected use of ethanol in fuel as oxygenate. Octane data showed wider range and layers reflecting three grades of gasoline available at pumps. [*See output of input no 567 in Jupyter notebook*]

Exploratory Data Analysis

Since this project was focused on Benzene, it's relationship with other constituents and properties were explored. Plots of every constituents and properties against Benzene in x-axis showed no obvious and/or compelling relationships [*Input no 570*]. The pearson's correlation coefficient between benzene and all other constituents/properties were calculated [*Output 575*]. The correlation coefficients were below 0.5 in both positive and negative direction. This indicated that the Benzene concentrations were independent of concentration of other constituents and values of other gasoline properties. Thus, it was more apt that Benzene concentrations be predicted by time series analysis.

The time series plot of Benzene concentration showed that data was available from 1975 to 2017[*Input 576*]. There was a gap between 1986 and 1987. The density of available data increased after 1995. The Benzene concentration decreased sharply after 1995 due to regulatory requirement. Benzene concentrations as high as 4.5% were recorded before 1995. After 1995, the maximum concentration remained below 1.5%. One value above 2.0% was observed but it was obviously an outlier. The mean Benzene concentration after 1995 was slightly above 0.5% [*output 529*].

The data variance/range can skew mean of the data sometimes. So, the data was resampled and monthly mean and median were compared [*Output 530*]. There was no obvious significant difference observed between the mean and median. The resampled monthly mean data series had some gaps. The gaps were filled with backfill method[*Output 531*].. The filled in data comprised of 26.5% filled values[*Output 578*]. The filled in data ranged from 0.43 % and 0.89 % with a mean of 0.62 %.

Results and Analysis

Three different modeling procedures were used to develop time series model and make predictions. In order to better generalize model, the data was split into train and test sets. The data between 1995 and 2010 was used for training models and data after 2010 was used for testing.

Simple exponential Smoothing

The simple exponential smoothing model was applied to the training data set with values of smoothing level parameter ranging from 0.1 to 0.9. The RMS increased almost linearly with increase in smoothing level [*output 534*]. The smoothing level of 0.1 produced a model with root mean square of 0.091. This model predicted an average Benzene concentration of 0.67 % through the test period of 2011 to 2017[*Output 536*].

Holt linear trend model

The Holt linear trend model extends the simple exponential smoothing by including an parameter for simulating trend. Thus, this model has smoothing parameter for level and trend. The smoothing level parameter of 0.1 derived from the simple exponential smoothing was kept. Other parameters were optimized by the Holt module of the statmodels api. This automatic optimization produced a model with root mean square of 0.12. The model predicted mean Benzene concentration of 0.71% with a positive trend during the test period of 2011 to 2017 [*Output 539*]. The optimized model parameters were shown [Output 540]. The model may be tweaked further to improve the RMS value. However, the for the purpose of this project, RMS of 0.12 is satisfactory.

Seasonal ARIMA model

Seasonal Auto Regressive (AR) Integrated (I) Moving Average (MA) models are more sophisticated than the exponential smoothing models employed earlier. These models are used to take in account of the trend as well as seasonality of the data. Stationarity of data must be checked before employing the model. The annual rolling mean and standard deviation plot show that the time series not stationary [*Output 541*]. The dickey-fuller test was also performed to test the stationarity of the time series [*Output 542*]. The test shows that the test statistic value is higher than even the 10% of the critical value. This test affirms the non-stationarity of the data.

The trend, seasonality and residuals of the time series were shown in a decomposition plot [*Output 543*]. There was no clear trend but seasonality of the data was quite clear. The seasonal pattern repeated every 6 months. The auto-correlation function and partial auto correlation functions were also plotted [*Output 546*]. The acf plot showed almost a sinusoidal pattern with the autocorrelation decreasing to lowest value only at lag of 6. The pacf decreased rapidly.

The plots and tests indicate that the seasonal ARIMA is a good choice to model this time series. In order to estimate the parameters of the model, grid search approach was used. The range of p,d,q parameters of ARIMA and Seasonal parameters P.D,Q was restricted to range of 0 to 1. This was decision was primarily taken from the pacf plot. The seasonality of 6 was chosen as the decomposition plot and the acf plot showed this to be appropriate. AIC was used to measure model performance. All possible combination of the parameters were generated and data was fit to corresponding model. The AIC was calculated for each model. The AIC provides a comparative measurement for the goodness of fit of the model. Comparatively, lowest the AIC, better fitting is the model.

The lowest AIC was computed for the model with following combination of paremters [*Output 549*]:
(p,d,q)X(P,D,Q,S) = (1,1,1)X(1,0,1,6)

The data was fit to the SARIMAX model of the statmodels api with the aforementioned parameter combination. The model coefficients were all shown to be statistically significant and standard error for the coefficients were quite small [*Output 550*].

Diagnostics plots of the model revealed that the residuals of the model were normally distributed [*Output 551*]. The standardized residuals were randomly distributed with a mean of 0. The histogram and q-q plot also show that model upholds the normality assumption. The correlogram shows no significant correlations with lags from 1 through 10.

To validate the model, two different forecasts from the model were compared with the observed data: static forecast and dynamic forecast. The static forecast generates the forecast at each point using the full history upto that point. The static forecast from the model was very close to observed data and produced a RMS of 0.049 [*Output 553*]. Static forecast predicted the Benzene concentration to vary between a narrow range of 0.45 % and 0.85 %. Dynamic forecast generates predictions after 2011 using the model fit until 2010. So, this is serves as a more rigorous and independent forecast. The RMS of the dynamic forecast was 0.091 which is more than satisfactory as it is comparable to simple exponential smoothing RMS but takes in account of the seasonality. The mean Benzene concentration predicted by dynamic forecast was closer to 0.7 % [*Output 556*].

The model was then used to make a long time forecast, 50 months beyond the scope of the available data [*Output 559*]. The long term forecast shows the mean benzene concentration to be around 0.62. The confidence interval of the forecast can be seen widening as time progresses. However, the model performs well enough to predict the concentration within -.15/+.15% even after 50 month period.

Conclusion

The models show that the Benzene concentration has remained below 0.9 % after 1995 and closely fluctuated around 0.6% after 2011. This shows that the regulatory requirements have had a definite impact on the benzene concentration. The forecasts show that benzene concentrations in gasoline can be estimated to be 0.62±.15% by vol of gasoline as long as current regulatory requirements remain in place.