**Forecasting Beijing Air Quality Conditions**

*Tamara Monge*

June 13, 2018

1. **The Problem**

Beijing’s air quality has been an enormous health concern in recent decades as the city has become further industrialized and more and more of its citizens have begun driving automobiles. Over recent years, the city has improved its air quality significantly. However, those living in Beijing must remain vigilant to air quality forecasts. In this project I analyze patterns in Beijing’s PM 2.5 (particulate matter 2.5 microns or less in diameter) concentration and forecast conditions for the future.

1. **The Client**

The client in this study is the people of Beijing. PM 2.5 is especially dangerous because it can pass through the human body’s natural filters and enter the lungs. Health concerns related to PM 2.5 include heart and lung disease, asthma, bronchitis, and other respiratory problems. The people of Beijing will be interested in having a fast, accurate, and computationally inexpensive method for forecasting the PM 2.5 concentration.

1. **The Dataset & Wrangling Steps**

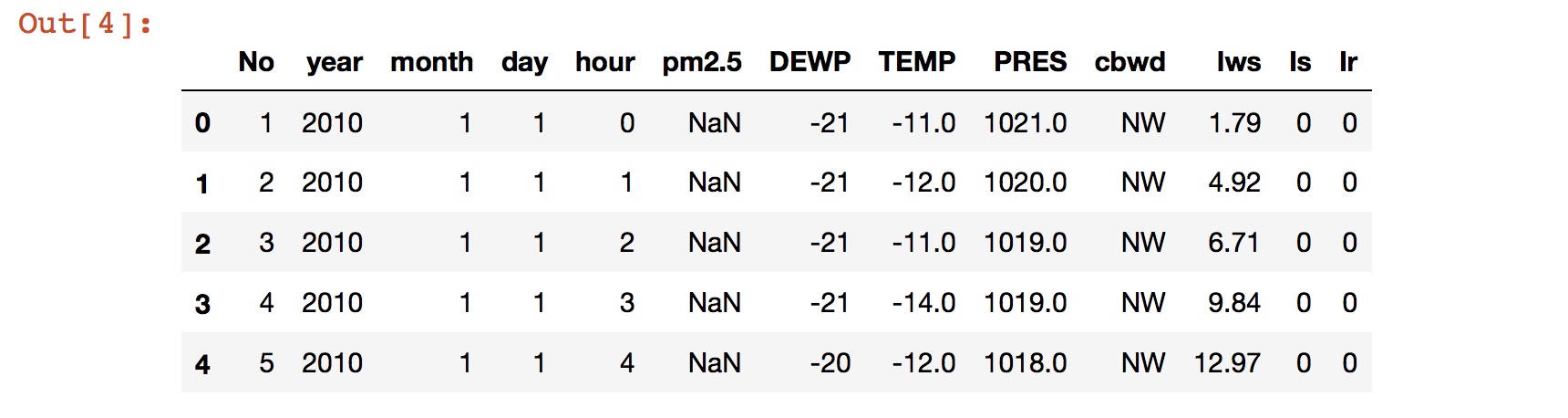
The dataset I am using for this project comes from UCI’s Machine Learning Repository: <http://archive.ics.uci.edu/ml/datasets/Beijing+PM2.5+Data>. It is available for free download and comes in CSV format.

The dataset consists of hourly observations between January 1, 2010 and December 31, 2014, for a total of 43,824 observations. Each observation has 12 features: year, month, day, hour, PM 2.5 concentration, dewpoint temperature, temperature, pressure, combined wind direction, cumulated wind speed, cumulated hours of snow, and cumulated hours of rain. The dataset is approximately 2 MB in size.

**3.1 Downloading and importing data**

I imported the dataset as a Pandas dataframe. The first five rows are shown in Table 1.

**Table 1. First five rows of raw dataset.**

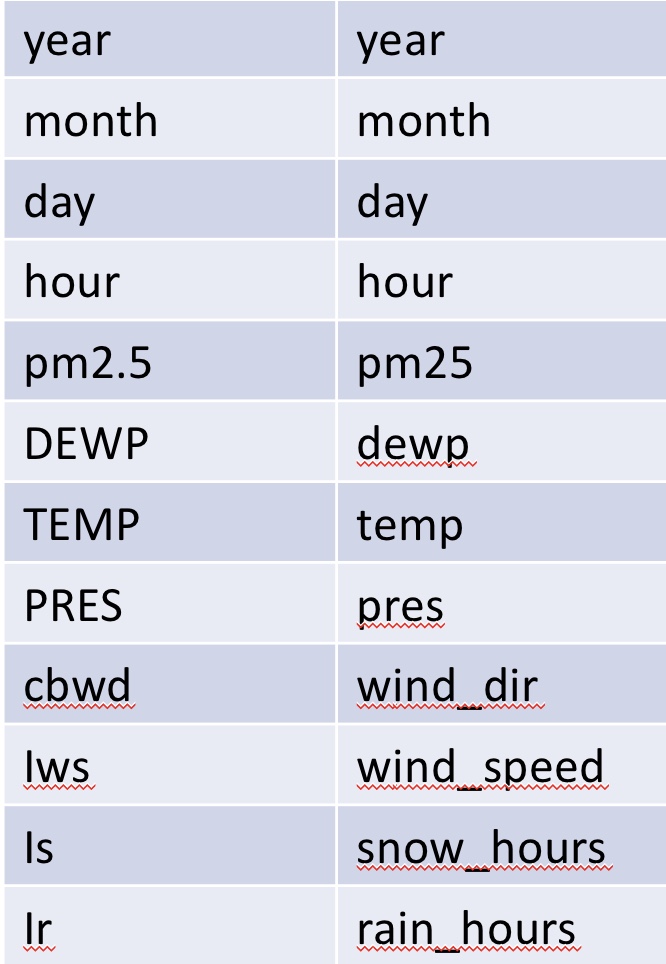


**3.2 Renaming columns**

In order to make data operations more convenient, I renamed some of the column titles. The original and renamed column titles are shown in Table 2.

**Table 2. Original and renamed column titles.**

**Original Renamed**



**3.3 Assigning datetime index**

I converted the year, month, day, and hour columns into a datetime object date. To assist with analysis and plotting, I then assigned date to be the datetime index of the dataframe.

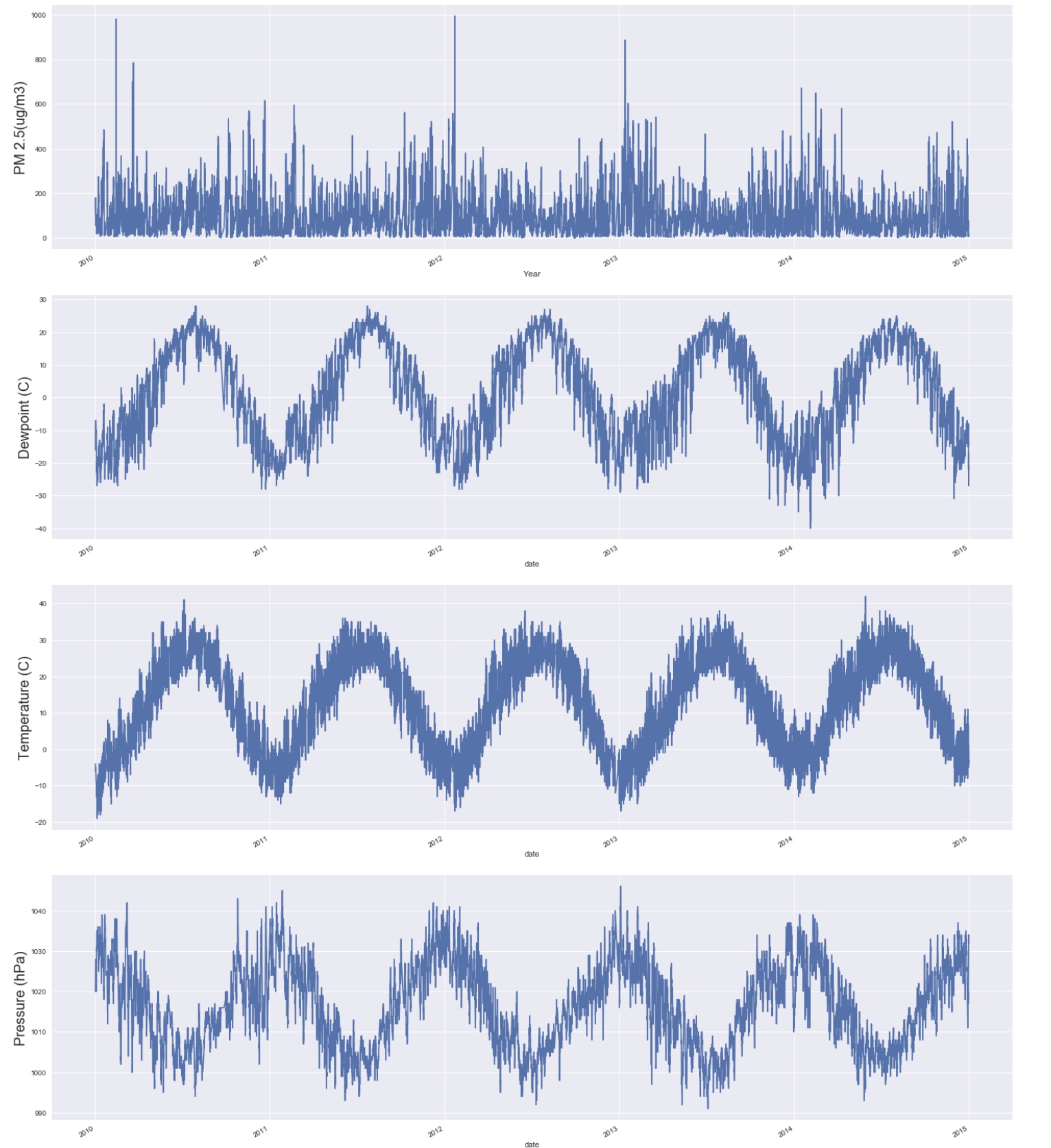
I also dropped the now redundant year, month, day, and hour columns.

**3.4 Dealing with missing values**

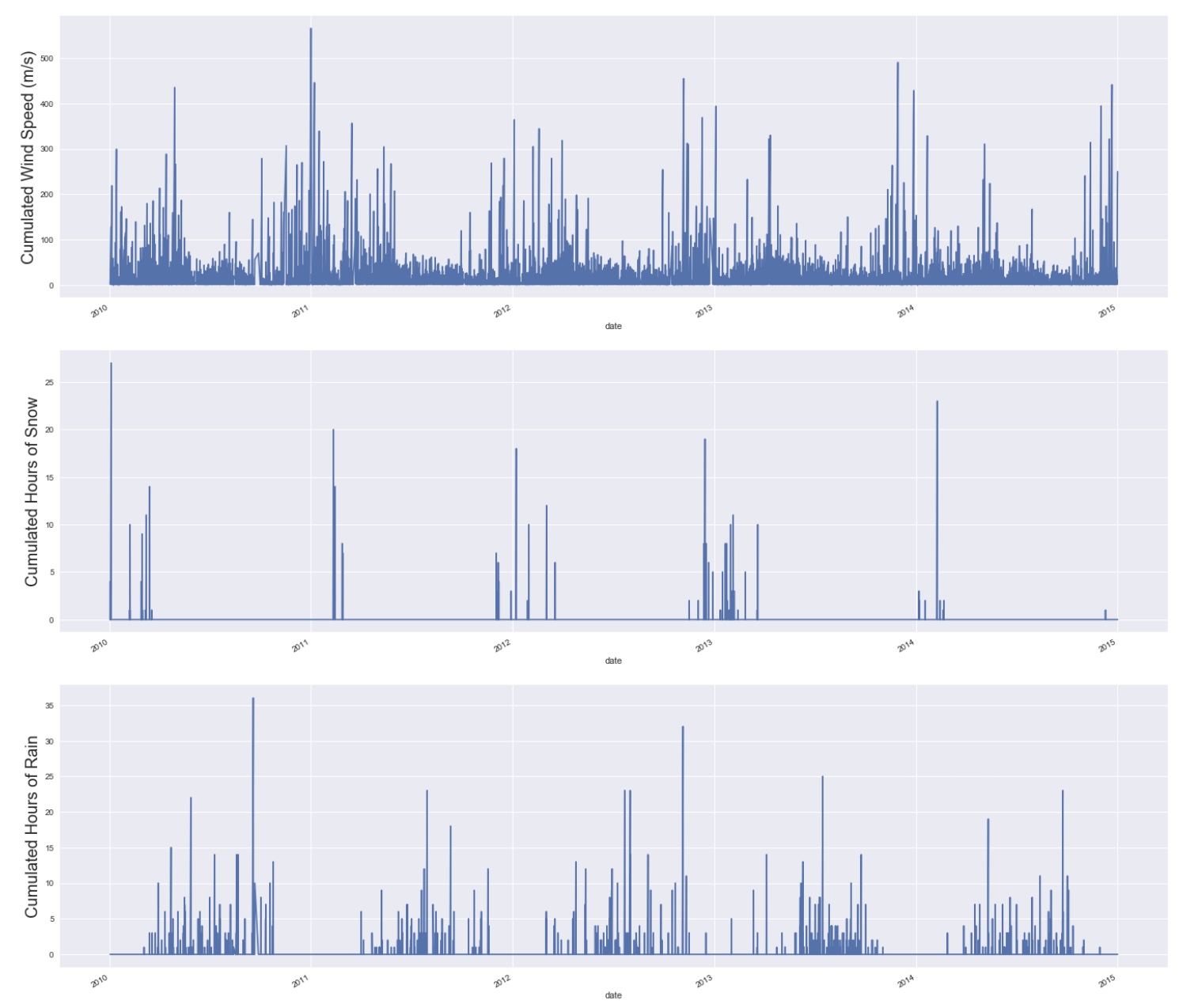
Approximately 4.7% of the observations were missing pm25 information. Since the missing data was such a small percentage, I decided it was reasonable to simply drop those observations. This brought the number of observations to 41,757.

1. **Exploratory Data Analysis**

The original time series for each feature are shown in Figures 1 and 2.



**Figure 1. Time Series of PM2.5, Dewpoint, Temperature, and Pressure.**

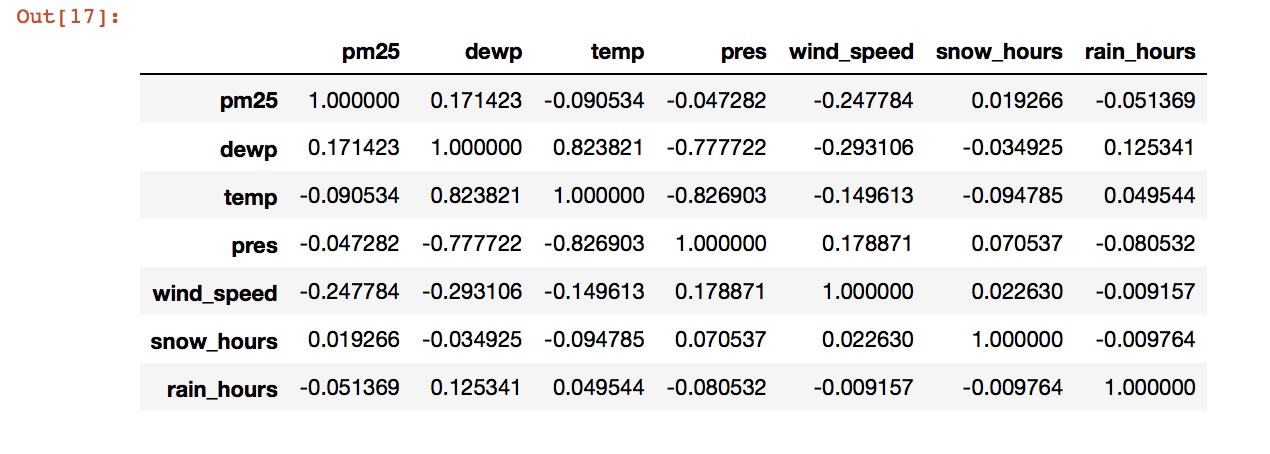
****

**Figure 2. Time Series of Cumulated Wind Speed, Cumulated Hours of Snow, and Cumulated Hours of Rain.**

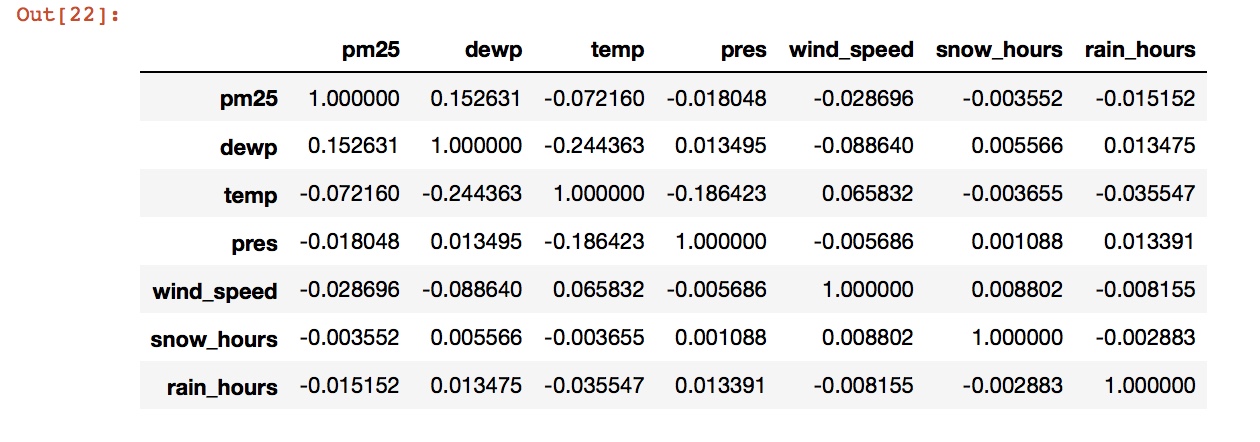
Visual inspection of Figures 1 and 2 reveals some interesting correlations between the observed fields. As expected, all seven features show strong seasonality. The seasonality of PM 2.5, pressure, wind speed, and hours of snow nearly align, while the seasonality of temperature, dewpoint, and hours of rain is offset from the others by ~6 months.

The correlation coefficients between the various time series are shown in Table 3. At first glance it appears that PM 2.5 is most strongly correlated with wind speed (-0.25) and dewpoint (0.17). However, we must keep in mind that these correlations combine seasonal and trend effects. The correlation coefficients after the trends have been removed (by way of first-order differencing) are shown in Table 4. There we see that the time series most strongly correlated with PM 2.5 is the dewpoint (0.15).

**Table 3. Correlation coefficients between original time series.**

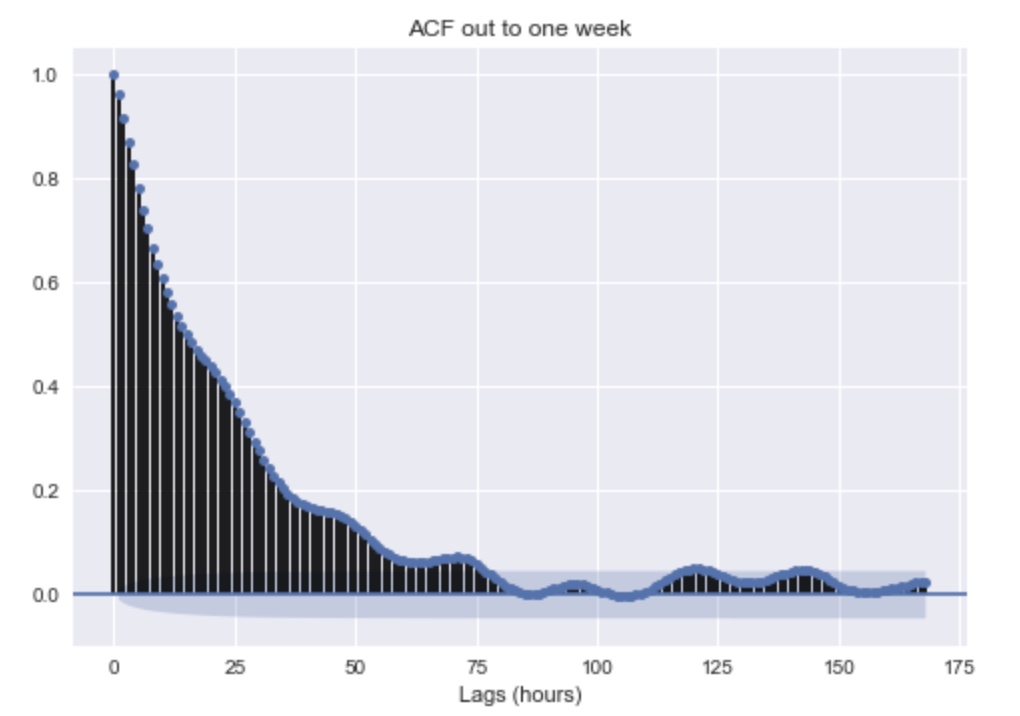


**Table 4. Correlation coefficients between differenced time series.**

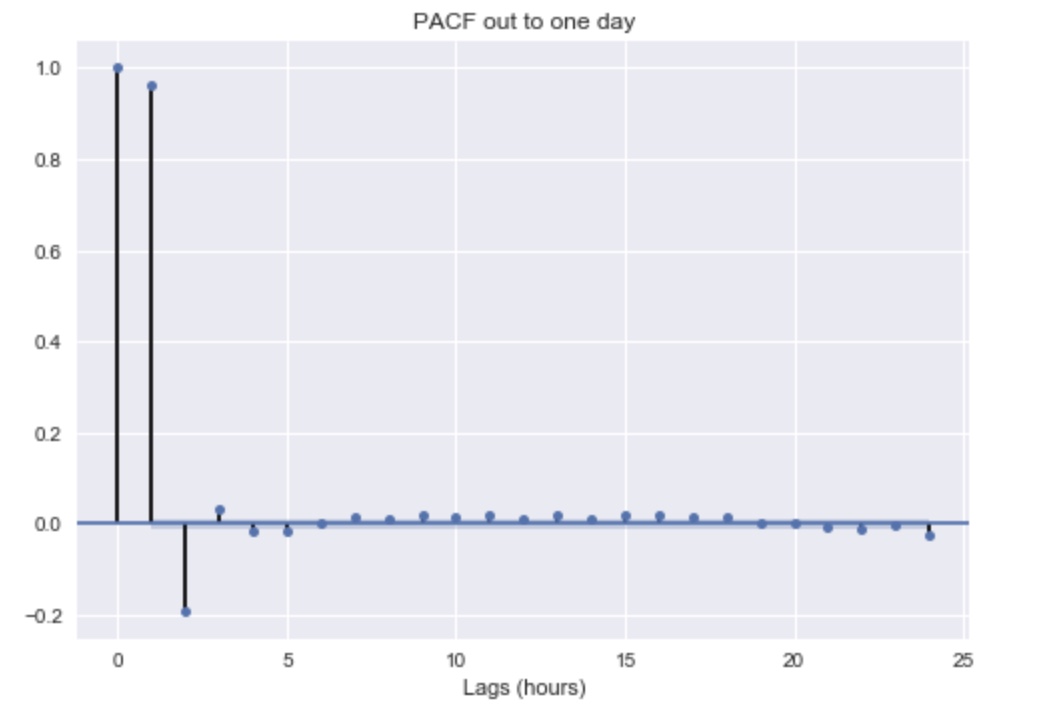


1. **Time Series Analysis**

First, let’s look at the autocorrelation and partial autocorrelation functions of the original PM 2.5 time series (Fig. 3-4).

****

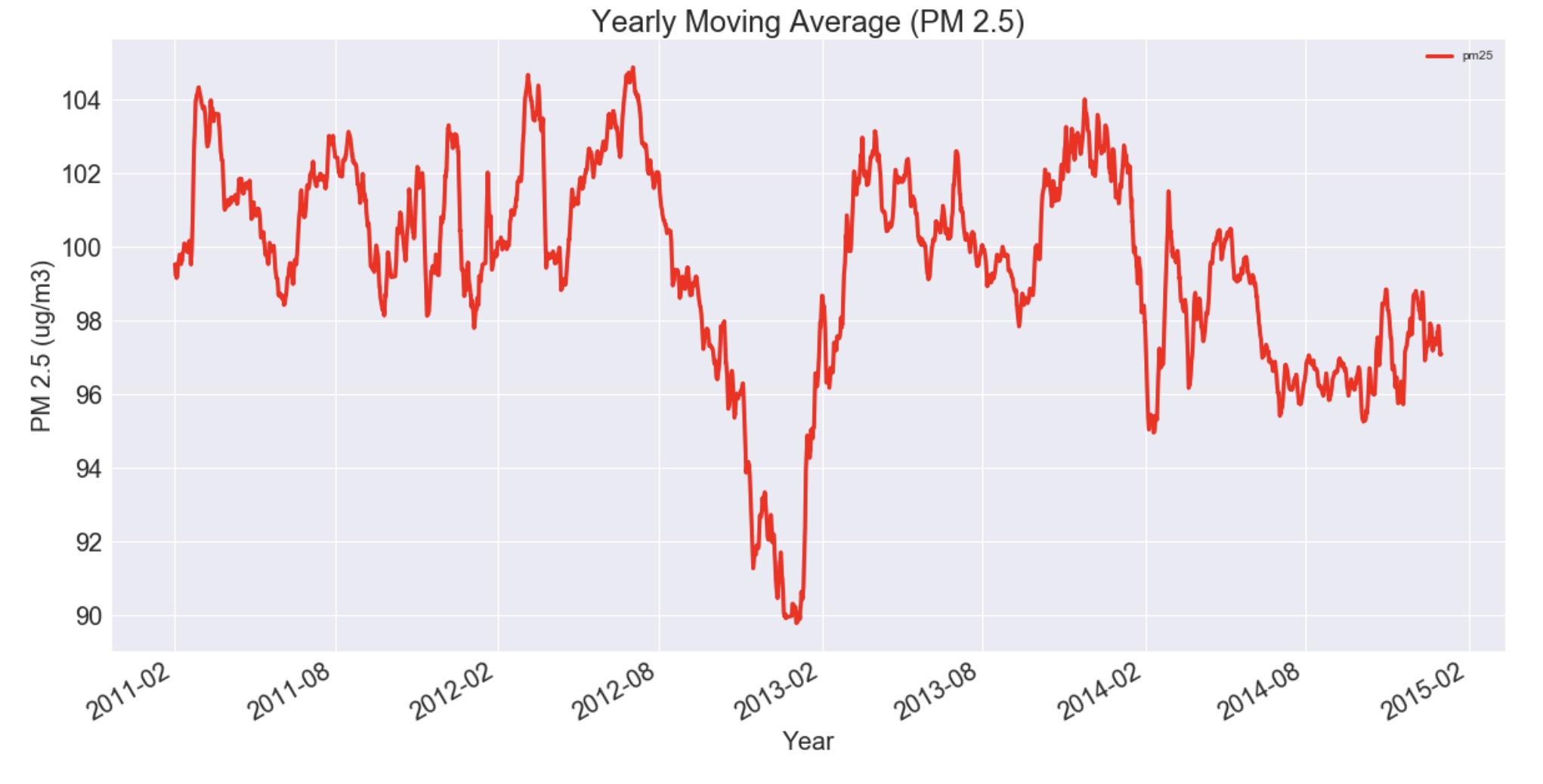
**Figure 3. Autocorrelation function of PM 2.5 time series.**

****

**Figure 4. Partial-autocorrelation function of PM 2.5 time series.**

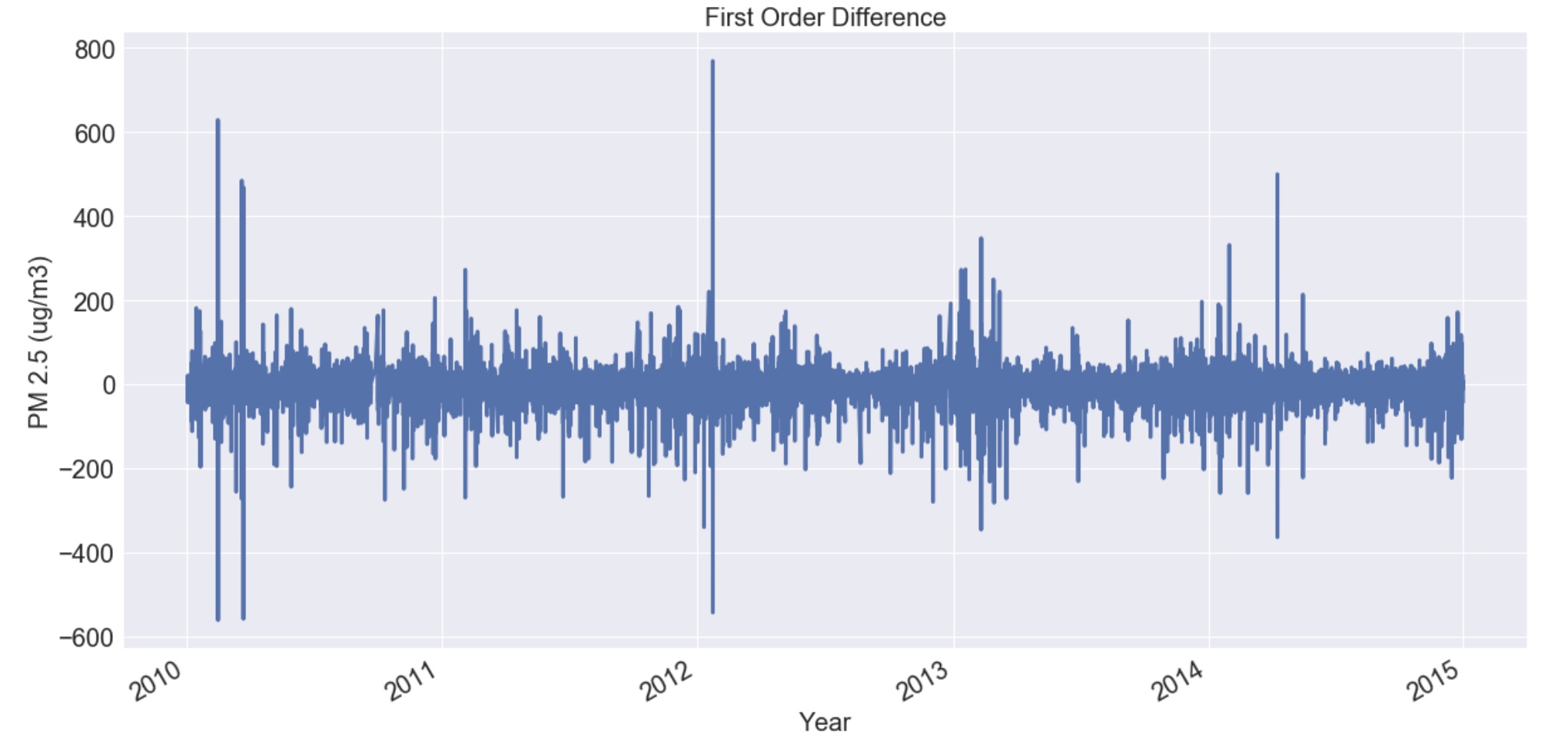
Figure 3 reveals a strong autocorrelation out to ~72 hours, while Figure 4 shows that most of that autocorrelation can be explained by the lags < 5 hours.

Now, let’s try to stationarize the time series. The moving average reveals the trend, so let’s look at that first (Fig. 5). There is a fairly steady trend between early-2011 to mid-2012, a sharp decreasing trend into early-2013 followed by a sharp increase into mid-2013, and a slowly decreasing trend into 2015.



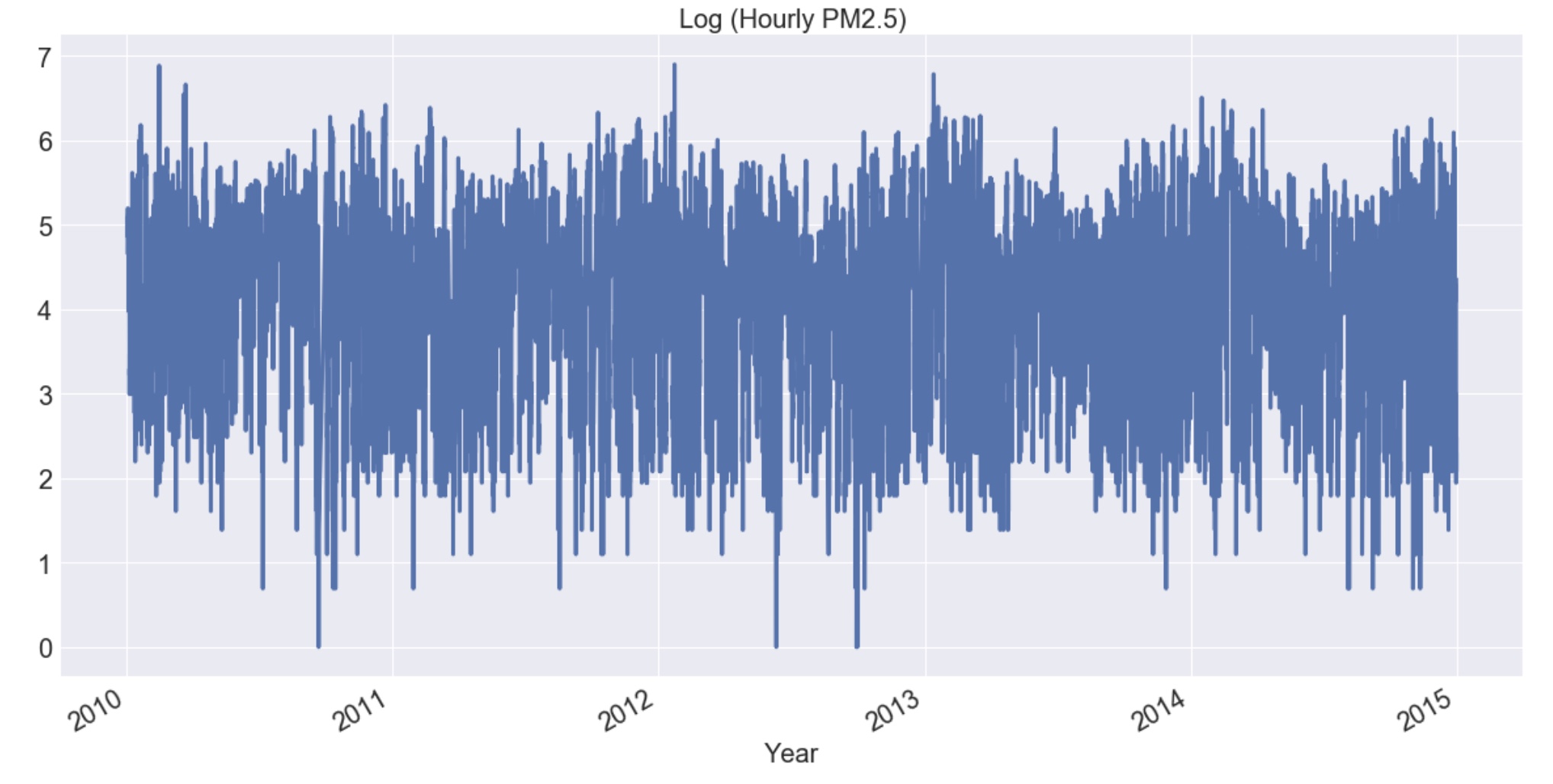
**Figure 5. Yearly moving average of PM 2.5.**

Now that we’ve identified the trend, let’s do a first-order difference to remove it (Fig. 6). Figure 6 shows that differencing removed the trend nicely, however there are still some large variations in the variance.



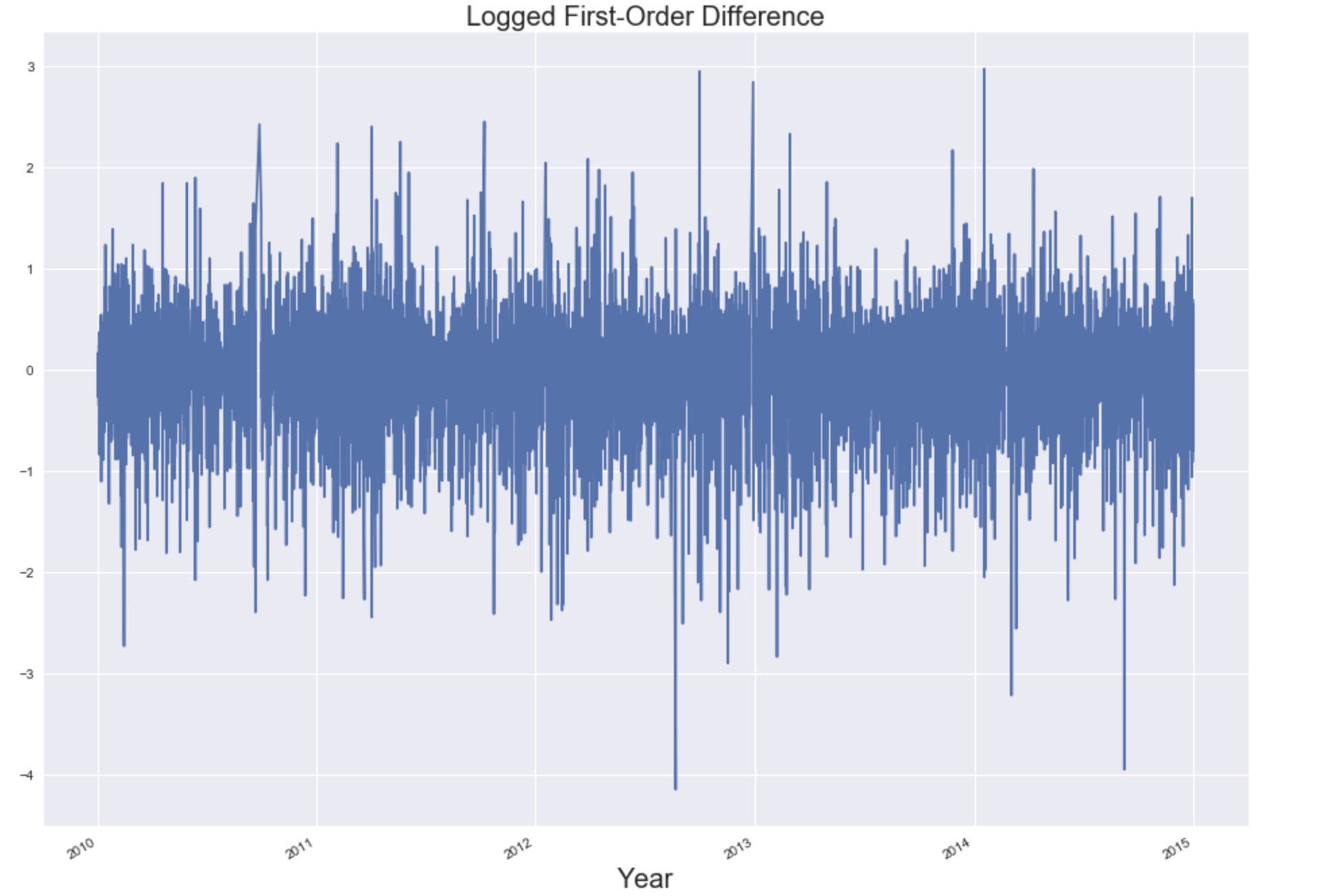
**Figure 6. First-order difference of PM 2.5.**

To stabilize the variance, let’s perform a log transformation (Fig. 7). Figure 7 shows that taking the log has made the variance relatively constant.



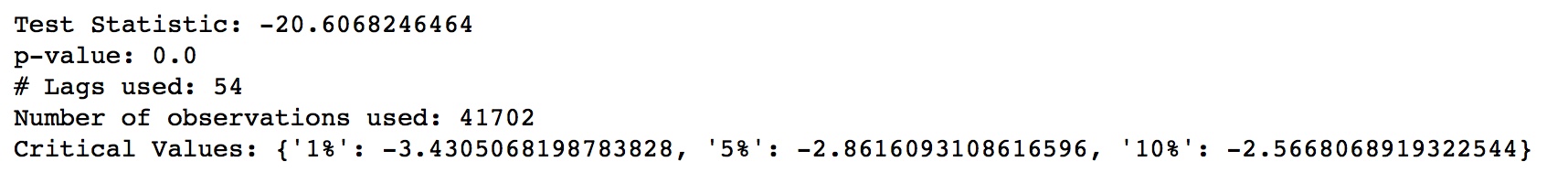
**Figure 7. Log transformation of PM 2.5.**

Combining these two transformations (Fig. 8) gives us a stationary time series model of the hourly PM 2.5 concentration.



**Figure 8. Logged First-Order Difference of PM 2.5.**

Now, we must test for stationarity statistically. To do this, we use the augmented Dickey-Fuller Test. The results of the test on the PM 2.5 time series are shown below:



According to these test results, the PM 2.5 time series is already stationary, even before applying the log and difference transformations.

1. **Next Steps**

The next step for this project is to fit various ARIMA models to forecast the PM 2.5 series. After that, I will stationarize the other fields’ time series and attempt to incorporate them into a forecasting model. Once the models are operational, the next step will be to evaluate model performance and create visualizations. Lastly, I will make recommendations to the client and offer ideas on how to extend the study.