Modeling the Air Pollution in China

– based on Gaussian Process

Zihan Zhang

July 27, 2021

The Gaussian Process (GP) is a powerful model in machine learning that is capable to represent the distribution of functions. In this report, I first go through some basic concepts and intuitions of how the Gaussian Process works, and then bring the model to the air pollution data in China. I apply the Gaussian Process to model the air pollution from the perspective of time and geography. The result shows that the pollution level experiences different hour trend between northern and southern cities, and northern cities may experience more sever pollution in aggregate.

1 Some basics about the Gaussian Process

The Gaussian Process (GP) is a powerful model that is capable to represent the distribution of functions. The vital idea of GP in modeling functions is by assuming a multi-variate Gauss distribution. Each input x can be regarded as a realization of a random variable X, and the correlation of those underlying random variables are assumed to be jointly normal distributed.

The multi-Gaussian distribution contains two key parameters, which are the mean and variance-covariance matrix. In GP those two parameters are shows as the functions, which are the mean functions $m(x)$ and the kernel functions $k(x, x')$. They are defined as follows.

$$\begin{aligned} m(\mathrm{x}) &= \mathbb{E}[f(\mathrm{x})] \\ k\left(\mathrm{x}, \mathrm{x}'\right) &= \mathbb{E}\left[\left(f(\mathrm{x}) - m(\mathrm{x})\right)\left(f\left(\mathrm{x}'\right) - m\left(\mathrm{x}'\right)\right)\right] \end{aligned} \tag{1}$$

Usually, we let the mean function $m(x) = 0$ for simplicity.

The kernel function $k(x, x')$ controls for the smoothness of the modeling function, since it describes the correlation of two inputs x and x'. A simple version of the kernel function is,

$$k\left(x, x'\right) = \exp\left(-\frac{\left(x - x'\right)^2}{2}\right) \tag{2}$$

, which implies the more correlated the x and x' are, the larger the value is. If x and x' are exactly the same, then the value of the kernel function goes up to one. Otherwise, the value are more closed to zero as x and x' are more divergent. Therefore, as the basic idea of GP, we can model a function $f(x)$ by the Gaussian Process, i.e.

$$f(\mathrm{x}) \sim \mathcal{GP}\left(m(\mathrm{x}), k\left(\mathrm{x}, \mathrm{x}'\right)\right) \tag{3}$$

In other word, the Gaussian Process modeling describes in (3) is actually modeling the function $f(x)$ based on the input x with a multi-variate Gaussian distribution. After that, the remaining issue is to apply the model for prediction based on some new input x^*. Also assuming the mean functions are zero, the extended GP with test data can be denoted as,

$$\begin{bmatrix} f(x) \\ f(x^*) \end{bmatrix} \sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} k(x,x) & k(x,x^*) \\ k(x^*,x) & k(x^*,x^*) \end{bmatrix}\right) \tag{4}$$

Equation (4) describes how do we modeling the function based on the observed data x, as well as, the prediction with test data x^*. However, in most cases, the observed data are not the $(x, f(x))$ exactly since we usually cannot avoid some noise. Therefore, it is common of introducing the noise term ϵ in to the function, i.e. make $y = f(x) + \epsilon$. As a result, the Gaussian Process modeling changes to,

$$\begin{bmatrix} y \\ f(x^*) \end{bmatrix} \sim \mathcal{N}\left(0, \begin{bmatrix} k(x,x) + \sigma_n^2 I & k(x,x^*) \\ k(x^*,x) & k(x^*,x^*) \end{bmatrix}\right) \tag{5}$$

, where σ_n^2 is the variance of i.i.d distributed noise term ϵ.

Apply the Bayes rule, we are able to get the formula for the Gaussian Process regression.

$$\begin{aligned} f(x^*) \mid x^*, x, y \sim \mathcal{N}\Big(& k(x^*,x)\left[k(x,x) + \sigma_u^2 I\right]^{-1} y, \\ & k(x^*,x^*) - k(x^*,x)\, k(x,x)^{-1} k(x,x^*)\Big) \end{aligned} \tag{6}$$

2 The air pollution data

My data for air pollution in China comes from the National Environmental Monitoring Center (CNEMC). They publish the air pollution conditions for over 300 Chinese cities in detail on their website, which is accessible to the public. The conditions for air quality contains various dimension including PM2.5, PM10, CO, SO2, etc. Here I mainly focus on the Air quality index (AQI), which is a comprehensive index for air quality. The larger the index is, the worse the air quality. In addition, in the daily published data from CNEMC, the AQI ranges from 0 to 500. The summary statistics results are shown in Table 1.

In this report, I have access to the air pollution data in the year of 2016. The data set contains over 300 cities, and I successfully merged 285 of them with their geography location. The data set also have very detailed data for various pollutants per hour in the whole year. For simplicity, I keep the AQI only, and generate the average AQI for all cities in my sample so that the air pollution can be modeled with respect to the geography location. The approximate location for those cities are shown in Figure 1. Additionally, I also select seven representative ones from the sample and keep full observation of AQI over the whole year. That contributes to identify what the pollution level performs within the day. For example, the AQI level in Jan 1, 2016 of those seven cities are shown in Figure 2. It is obvious that those cities may have different trends of pollution within a day.

Table 1: Descriptive Statistics

	(1) N	(2) mean	(3) sd	(4) min	(5) max
AQI per hour in 7 representative cities					
Beijing	8,497	102.0	82.36	9	500
Shijiazhuang	8,560	132.7	102.2	15	500
Shanghai	8,559	65.88	41.98	9	452
Xiamen	8,558	46.31	23.04	7	221
Guangzhou	8,559	54.63	26.73	9	264
Chengdu	8,559	89.98	48.91	14	310
Urumqi	8,559	107.9	88.55	9	500
Average AQI over year for 285 cities					
AQI	285	74.22	20.42	29.15	132.7

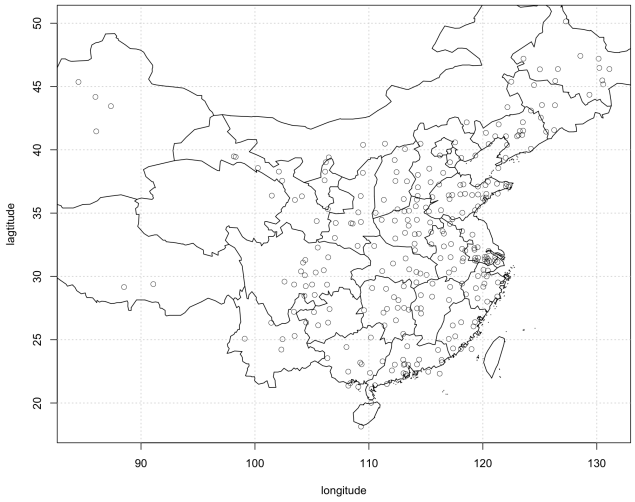


Figure 1: Approximate city location

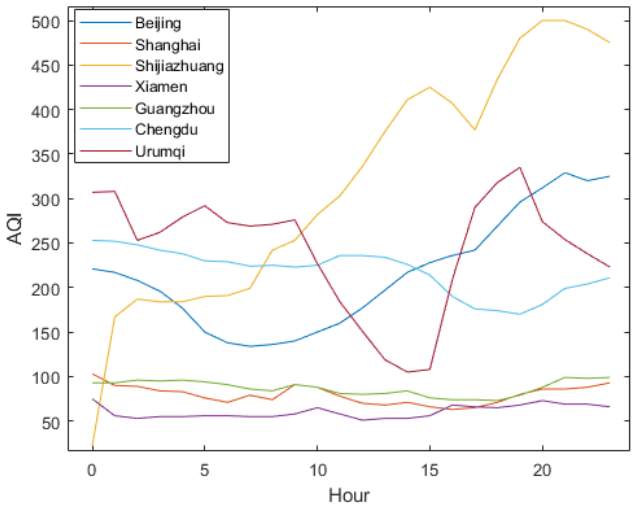


Figure 2: AQI on Jan. 1, 2016

3 Modeling the hourly time trend

3.1 Estimation

First, I bring the data of seven representative cities to the model described as equation (5) and (6). I intend to model the trend of AQI level within a day, therefore, the function form can be represented as

$$y = f(t) + \epsilon \tag{7}$$

, where t is the hour ranging from 0 to 23 in a day and y represents the AQI level.

Before training the model, I have to determine the form of kernel and likelihood functions. I apply the squared exponential (SE) covariance as the kernel function, which is one of the mostly used kernel function. The form can be denoted as,

$$k\left(t, t'\right) = \sigma_f^2 \exp\left(-\left(t - t'\right)^2 / \left(2l^2\right)\right) \tag{8}$$

, where σ_f governs the magnitude of overall variation, and l controls the the smoothness of $f(t)$.

With the kernel function, the likelihood function is corresponding as,

$$\log p(y \mid t, \boldsymbol{\theta}) = -\frac{1}{2} y^T K_y(\boldsymbol{\theta})^{-1} y - \frac{1}{2} \log |K_y(\boldsymbol{\theta})| - \frac{n}{2} \log 2\pi \tag{9}$$

, where $\boldsymbol{\theta}$ denotes the hyper parameter within the function, and $K_y(\boldsymbol{\theta})$ is denoted as,

$$K_y(\boldsymbol{\theta}) = k(t, t') + \sigma_n^2 I. \tag{10}$$

, with σ_n governs the variety of noise.

From the form of kernel and likelihood function, it is intuitive to recognize that the hyper parameter $\boldsymbol{\theta}$ contains three items, including σ_f, σ_n and l. Therefore, we are able to determine the value of them by maximizing the value of the likelihood function, i.e.

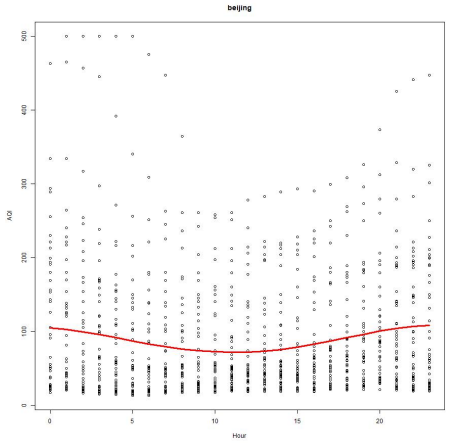
$$\hat{\boldsymbol{\theta}} = \arg\max_{\theta} \log p(y \mid t, \boldsymbol{\theta}) \tag{11}$$

3.2 Results

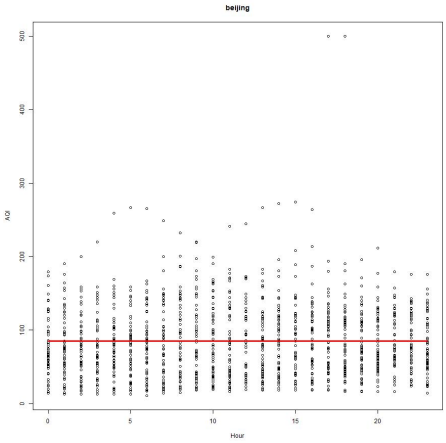
I take the GP model mentioned in previous sections to the hourly air pollution data for seven representative cities, and estimate the hourly trend of AQI for cities respectively. The results give the information that the trend may probably different among cities. Additionally, considering that the hourly pollution may performs differently in seasons, I also separate the data according to seasons and estimate the model with separable data so that it is able to capture the seasonal features for pollution trends.

The estimation results are shown in Table 3. The left panels are the AQI trends in winters while the right panels displays the summer ones. Each row represents a particular city, with the first 3 rows displaying cities in north China and the last four rows showing those locate in south China. The red curve in each panel describes the estimated trend by Gaussian Process method based on the particular data that drawn as points in each figure.

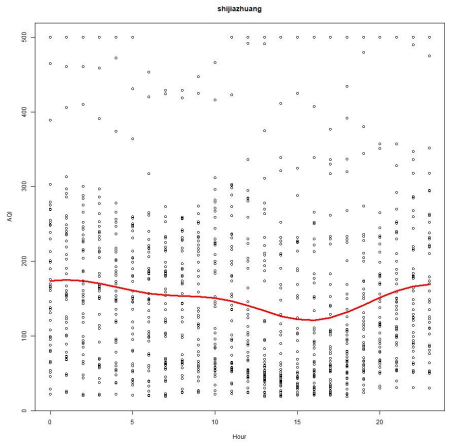
The figures display significant and distinguishable results that are completely different seasonally and geographically. In northern representative cities shown in panel (a) to (f), including Beijing, Shijiazhuang and Urumqi, the hourly trend in winter reflects an U shape from 0 to 23. This indicates that the pollution are more severe at night. In winter, the AQI of northern cities gradually decreases from mid night till about mid-day, then increases slowly. However, in summer, it gives a completely different hour trend that both Beijing and Shijiazhuang do not experience significant hourly change, except for small fluctuation in Urumqi.



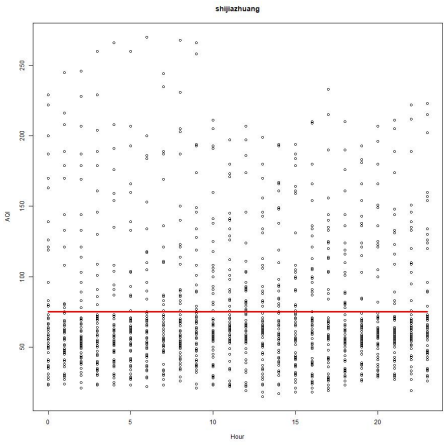
(a) Beijing Winter



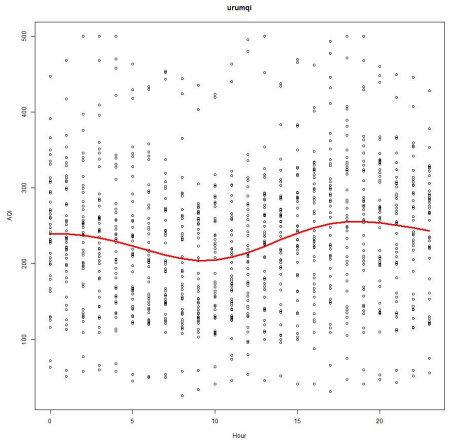
(b) Beijing Summer



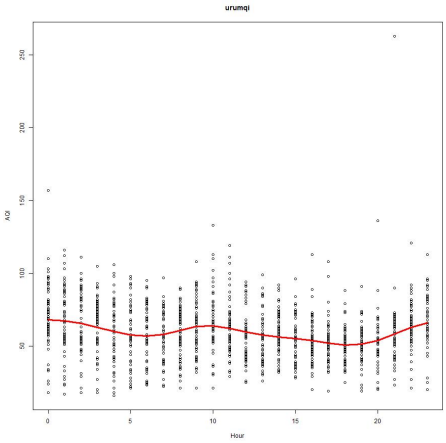
(c) Shijiazhuang Winter



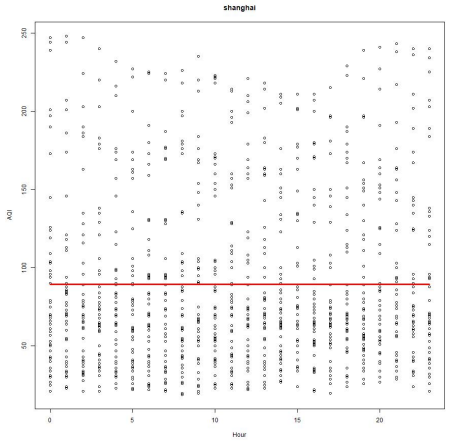
(d) Shijiazhuang Summer



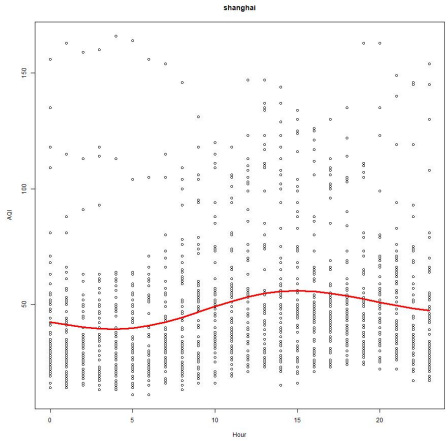
(e) Urumqi Winter



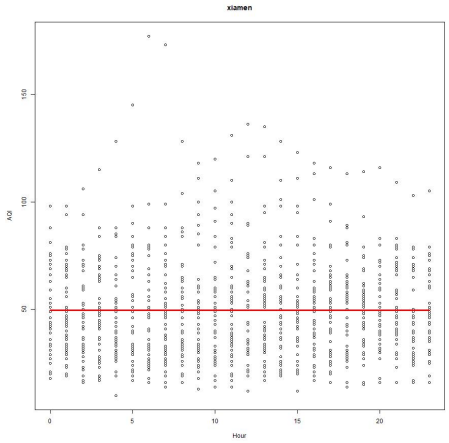
(f) Urumqi Summer



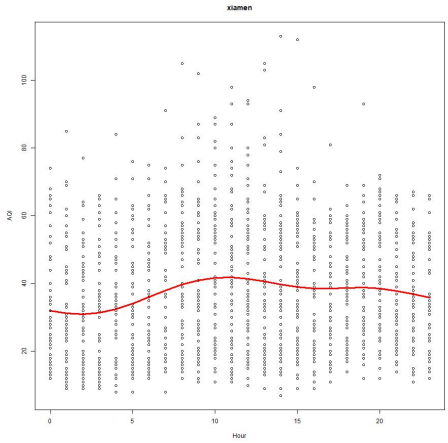
(g) Shanghai Winter



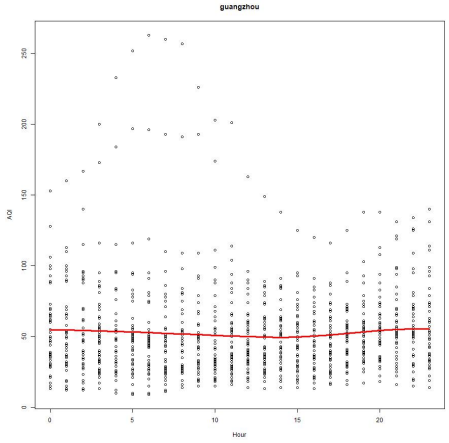
(h) Shanghai Summer



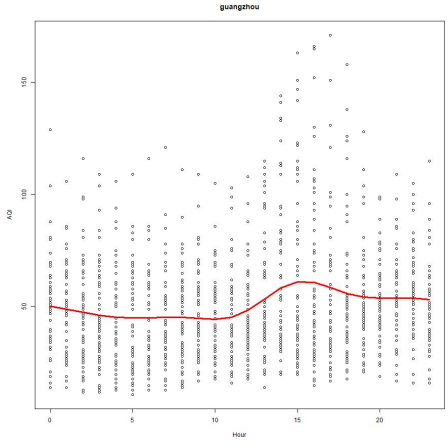
(i) Xiamen Winter



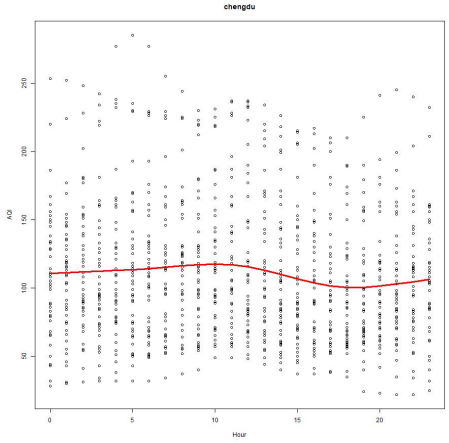
(j) Xiamen Summer



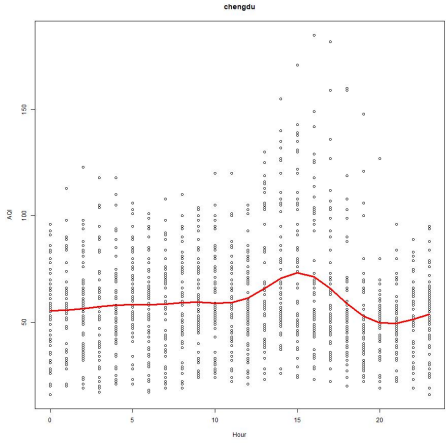
(k) Guangzhou Winter



(l) Guangzhou Summer



(m) Chengdu Winter



(n) Chengdu Summer

Figure 3: AQI hourly trend

The trend in southern cities shown in panel (g) to (n) are quite different seasonally. In winter, the trend are quite similar as that of northern cities in summer, which reflects a flat predicted hour trend. However, in summer, the trend in southern cities shows an inverse U-shape. That indicates those cities may have less pollutants at night comparing to daytime.

4 Modeling the geographical distribution

4.1 Estimation

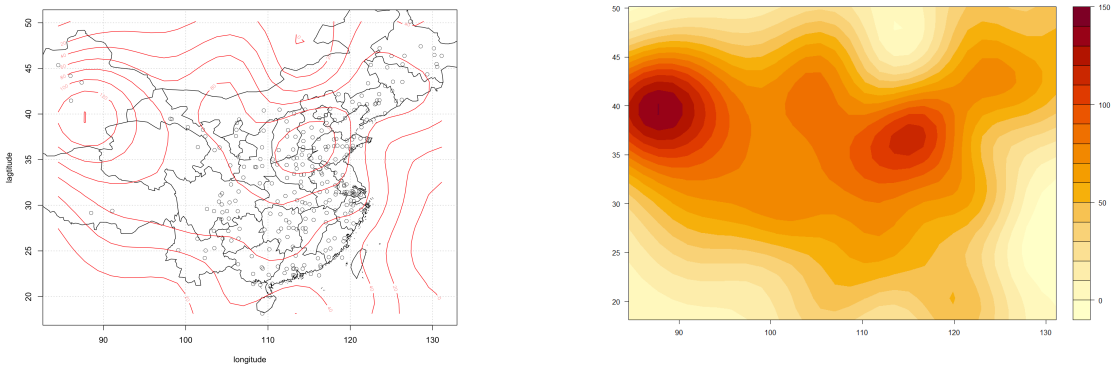
The only difference of modeling the geographical feature of AQI distribution from modeling time trend is the dimension of inputs. It requires to build up a model based on both latitude and longitude. Hence, there are some small modification about the kernel function when applying Gaussian process with multiple inputs. One natural way for the kernel function is simply add the kernel for different dimensions together.

$$\begin{aligned} K &= k_1(x_1, x_1) + k_2(x_2, x_2) + \cdots + K_d(X_d, X_d) \\ K_y &= K + \sigma_n^2 I \end{aligned} \tag{12}$$

Here d represents the dimensions of inputs. While the other parts are the same as what I have done in section 3.1.

4.2 Results

Here I take the model to daily average AQI data in 2016 to capture the geographical distribution of air pollution level. The results are displayed in Figure 4. Both panel (a) and (b) describes the estimated AQI level based on the longitude and latitude. In panel (a) I showed the contour curve on the map with city location shown as points, and panel (b) shows the estimated AQI level in a more intuitive way.



(a) China map with estimated AQI contour

(b) Estimated AQI heat map

Figure 4: Estimated AQI distribution in geography

The results indicates that there are two regions suffers from severe air pollution relatively. The one is Huabei area in north China, and the other one is Xinjiang. Moreover, the coastal regions may have better air quality over year comparing to the inland regions, and the southerns look better than the northerns. Those results are quite intuitive because it is common that

northwestern regions in China including Xinjiang are more probable to suffer from sandstorm coming from overseas, and Beijing and adjacent regions are often well-known by the heavy industries such as iron industry. In addition, the coastal regions may have more rains, which contributes to the reduction of pollutants. These could be reasonable explanations for the estimated results.

5 Conclusion

In this report, I first go through some basic concepts and key procedures of building up a Gaussian process model, which is a powerful model that is able to capture the function distribution. Next, I take the model to some real data of pollution in China. The Gaussian process can be useful in modeling the hour trend and the distribution of air pollution, which are two cases requiring only single and multiple inputs respectively. It also gives some interesting results including the seasonal and geographical difference of hourly trends, and also the average pollution level distribution on average over year. Those evidences may reflect that northern cities in China may have most severe pollution at night within a day in winter, while the southern cities are most pollutant at afternoon in summer. Additionally, I am able to discover the significant seasonal changes for this trend. Finally, the Gaussian process also contributes to point out areas with more serious air pollution.

References

[1] Carl Edward Rasmussen and Christopher K. I. Williams, *Gaussian Processes for Machine Learning*, The MIT Press, 2006. ISBN 0-262-18253-X. `http://www.gaussianprocess.org/gpml/`.

[2] Ximing Wu, *Lecture Notes on Gaussian Process methods and Applications.*

[3] Wikipedia, *Gaussian Process*, `https://en.m.wikipedia.org/wiki/Gaussian_process`.