Association between PM2.5 concentration and other meteorological factors

Qiwen Hua

4/19/2022

Introduction

PM2.5, also known as fine particulate matter, is a form of air pollutant with a size of 2.5 microns or smaller. It is well accepted that a high PM2.5 concentration imposes great concerns to people's health. Many cities in the world suffer from consistently high PM2.5 concentration, especially in Beijing, China. Therefore, in this analysis, we will use a dataset of Beijing PM2.5 and other meteorological information to analyse the association between PM2.5 concentration and other measurable factors. In particular, the factors of interest are hour of the day, dew point, temperature, and wind speed.

Methods

Data source

The dataset that we will use throughout this analysis comes from the UCI Machine Learning Repository gathered by Liang, X. et al (2015). The dataset contains hourly meteorological data from 2010 Jan 1st to 2014 Dec 31st of US Embassy in Beijing, which includes information such as the PM2.5 concentration, dew point, temperature, wind speed, etc. Among those features, listed below are our interested variables and their corresponding units:

1. PM2.5 concentration (μg/m^3);
2. Hour of the day;
3. Dew point (°C);
4. Temperature (°C);
5. Wind speed (m/s).

Data cleaning and wrangling

To answer the overall question raised in Introduction, we only need a subset of the features provided in the dataset. Namely, we need date time, PM2.5 concentration, dew point, temperature, and wind speed. Therefore, we will first use `lubridate` to create a new `datetime` variable from the `year`, `month`, `day`, and `hour` columns of the dataset, and use `dplyr` to rename and select the interested columns.

The dataset contains some missing values, denoted by `NA` in the imported dataset in R. Since the dataset is already large enough, we directly remove rows containing missing values instead of imputing to avoid introducing bias. We will use `tidyr` to remove such rows.

We have now finished cleaning the data. However, to make comparisons easier in later analysis, we will create some categorical variables for some variables. We will set the categories to be `high` to values that are 1 sd above the mean, `low` to values 1 sd below the mean, and `normal` to everything else.

In addition, we can categorize time of the day with the following rules:

1. `morning` - 00:00 to 11:59;
2. `afternoon` - 12:00 to 17:59;

Table 1: Summary statistics for interested variables

	Min	Max	Mean	SD	Low hours	Normal hours	High hours
Dew Point	-40.00	28.00	1.75	14.43	8576	24217	8964
Temperature	-19.00	42.00	12.40	12.18	9328	24346	8083
Windspeed	0.45	565.49	23.87	49.62	0	38326	3431

Table 2: Summary statistics for PM2.5 concentrations

Min	Max	Good hours	Moderate hours	Unhealthy hours	Very unhealthy hours	Hazardous hours
0	994	15945	10445	10139	3469	1759

3. `evening` - 18:00 to 23:59.

Lastly, following the PM2.5 levels definition posted by Blissair, we define air quality levels base on PM2.5 concentration with the following rules and add them as a categorical variable in the dataset:

1. `good` - 0 to 50
2. `moderate` - 51 to 100
3. `unhealthy` - 101 to 200
4. `very unhealthy` - 201 to 300
5. `hazardous` - 301 and above

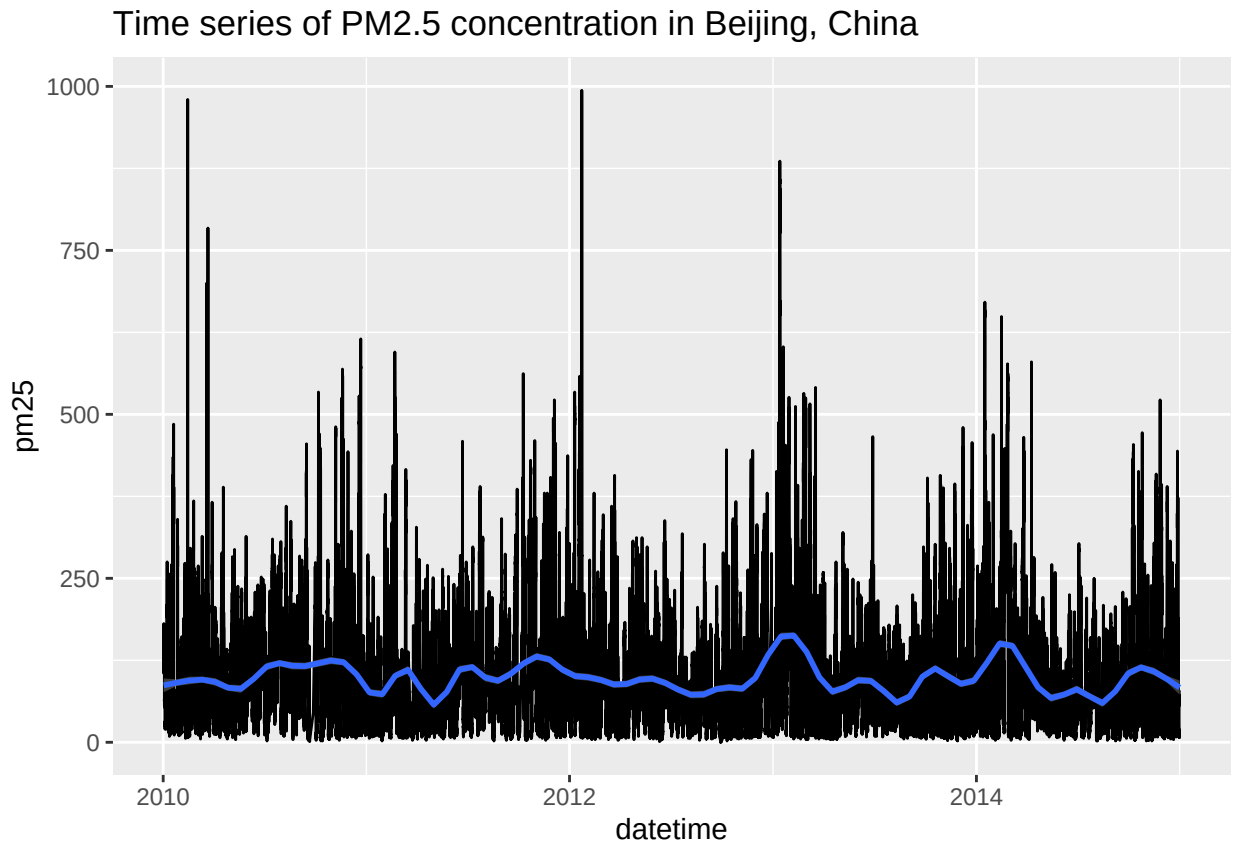
Data exploration tools

With the data cleaning and wrangling process finished, we can start conducting analysis with the processed dataset. In the next section, we will try to answer the goal of this analysis, finding out the association between PM2.5 concentration and some measurable factors, with methods such as creating figures with `ggplot2`, group and summarise with `dplyr`, and constructing generalized linear models with the `mgcv` implementation of `gam`. Lastly, we will build decision trees to predict PM2.5 concentration levels using the `rpart` library.

Results

Overall trend of PM2.5 concentration

First, we can take a look at the time series of Beijing PM2.5 concentration over the five years (2010 to 2014, inclusive). This gives us an overview of annual and seasonal trends of PM2.5 concentrations. Since the variance is quite high, we will add a smooth line generated from a cubic regression spline bases with 40 knots (represented by the blue line in the figure below).



The graph displays no obvious seasonal nor annual trends, suggesting that the PM2.5 concentration may not be associated with season or year.

Summary of PM2.5 concentration grouped by various factors

To give us an overview on the association between PM2.5 concentration and our factors of interest (e.g. wind speed), we can take a look at the means and standard deviations of PM2.5 concentrations grouped by our factors of interest, i.e., part of the day, dew point, temperature, and wind speed.

Table 3: PM2.5 concentration by part of the day

day_part	PM2.5 mean	PM2.5 sd
morning	100.59975	91.75683
afternoon	87.47384	85.65362
evening	106.83448	98.75993

Grouped by part of the day To the contrary of many people's intuition, PM2.5 concentrations is the lowest during the afternoon and highest during the morning and evening. This suggests that the air quality is worse at night. We will try to further varify this later using a generalized linear model.

Table 4: PM2.5 concentration by temperature category

temp_cat	PM2.5 mean	PM2.5 sd
low	111.74464	116.26408
normal	96.82083	88.46526
high	88.85785	65.57329

Grouped by temperature The summary statistics suggests that the PM2.5 decreases as temperature increases. This may be due to the fact that low temperatures traps small particles (PM2.5) more than high temperatures which cause the air to move upwards. This result hints a possible explanation to the observation above: the temperature in the afternoon are usually the highest which yields a lower PM2.5 concentration.

Table 5: PM2.5 concentration by dew point category

dew_pt_cat	PM2.5 mean	PM2.5 sd
low	47.24743	58.37452
normal	113.63934	103.20946
high	107.16131	65.01520

Grouped by dew point Dew point values are the temperatures the air needs to be cooled to in order to have full relative humidity, i.e., higher dew points means higher amount of moisture in the air. The summary statistics here shows that the PM2.5 concentration is much lower during hours with low moisture in the air.

Table 6: PM2.5 concentration by wind speed category

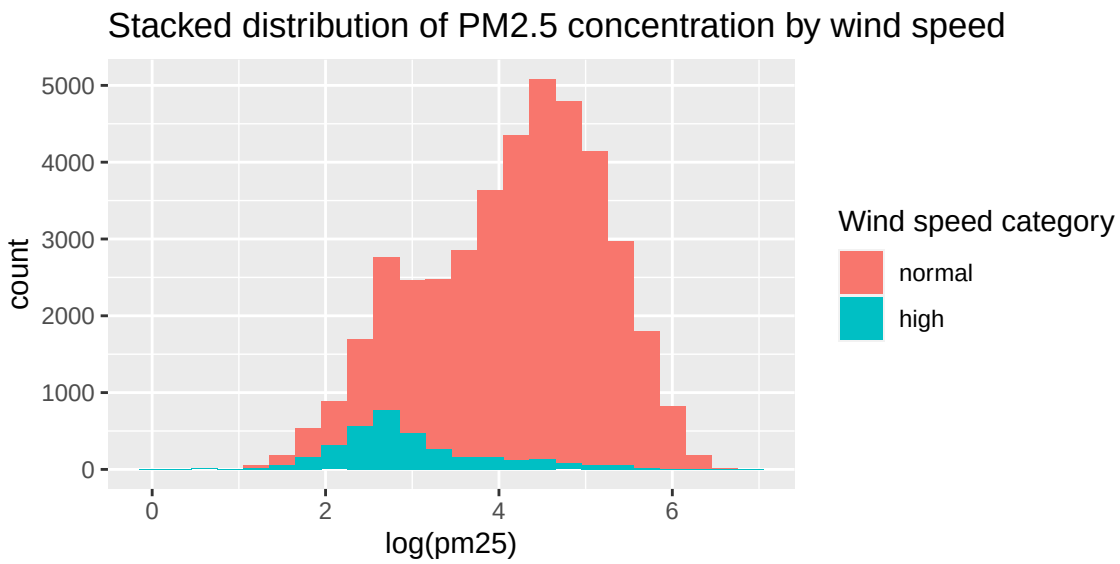
wind_spd_cat	PM2.5 mean	PM2.5 sd
normal	104.59435	92.89359
high	31.80093	43.22486

Grouped by wind speed Note that in section *Data cleaning and wrangling*, we fond that there are no hour in our dataset falls into the "low" wind speed category, i.e., no hour has a wind speed that is 1 sd below the wind speed mean. Therefore, our summary table here only has two rows.

From the table, we can see that hours with high wind speeds have much lower PM2.5 concentration than the hours with normal wind speed. This suggests that windy weather is associated with better air qualities, which may be due to wind being able to blow air pollutants higher up in the atmosphere.

Stacked distribution of PM2.5 concentration by wind speed category

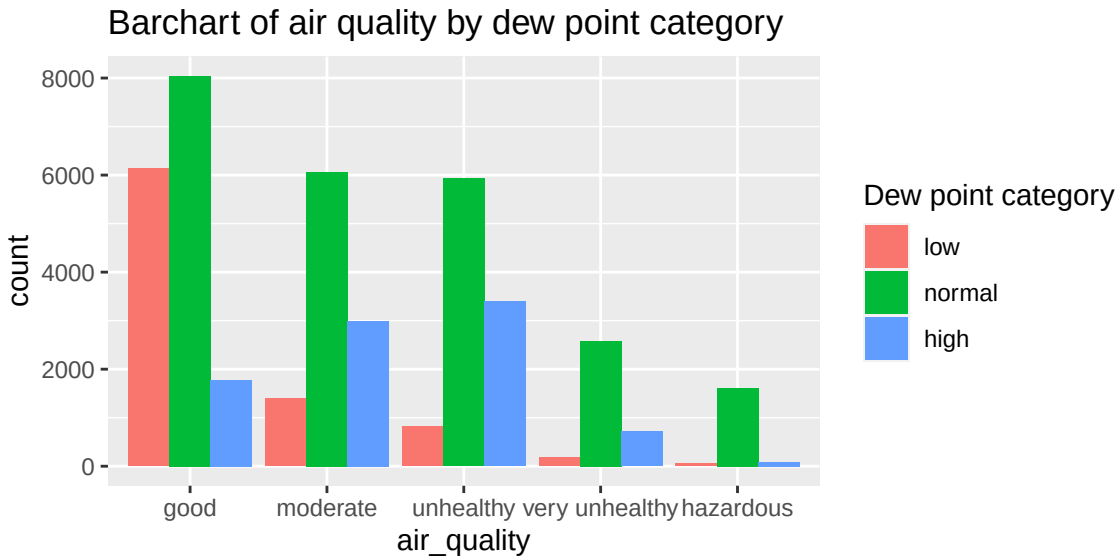
Following from previous results, we will look into the association between PM2.5 concentration and wind speed deeper. First, we will visualize the distributions of PM2.5 concentrations stacked using the two wind speed categories. Note that since PM2.5 concentration is not normally distributed, we will instead plot the log of PM2.5 distribution.



From the stacked histogram, we can see that the mode of the high wind speed PM2.5 concentration distribution is much smaller than the one for normal wind speed. This further suggests that high wind speed is associated with better air quality.

Barchat of air quality by dew point

Now we will investigate the assocaition between PM2.5 concentration (represented by air quality categoies) and dew points. We can visualize the association by plotting side-by-side barchats of air quality colored by dew point category.



From the plot above, we notice that the proportions of low dew point hours are much higher in good and moderate air quality hours. Combining from previous results in section *Summary of PM2.5 concentration grouped by various factors*, the dataset suggests that low dew points is associated with lower PM2.5 concentrations.

Generative additive model of hourly PM2.5 mean

Now we want to further investigate the association between the hour of the day and POM2.5 concentration. Previous results obtained in section *Summary of PM2.5 concentration grouped by various factors* shows that

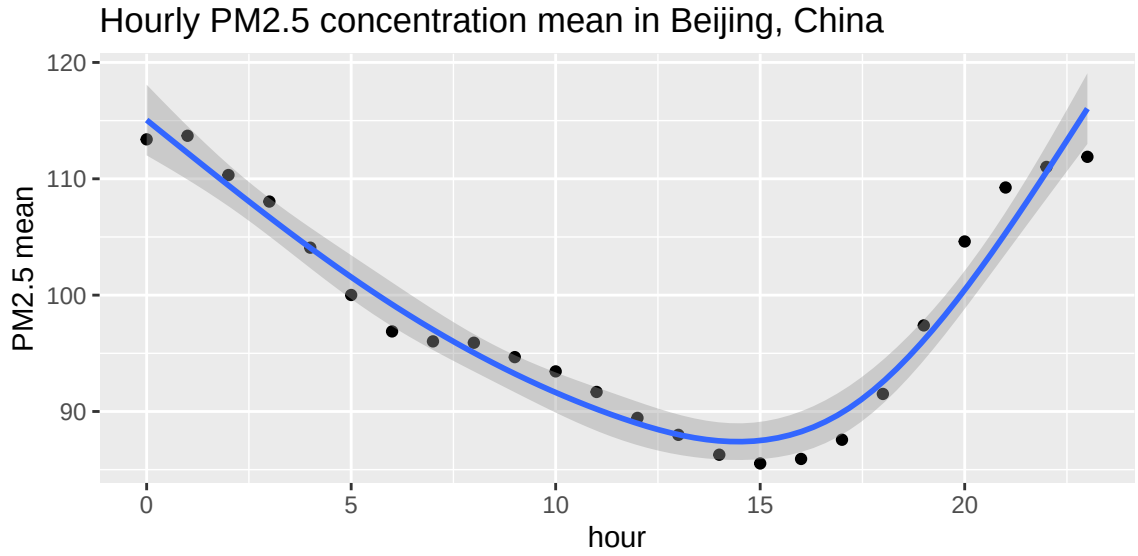
afternoon hours have the relatively low PM2.5 concentrations, and we want to verify that by visualizations and regressions.

First, we will create a new dataset consisting hourly PM2.5 concentration means of the entire dataset. Here are the first 3 rows of the new dataset.

Table 7: First 3 rows of the hourly PM2.5 mean dataset

hour	mean
0	113.3902
1	113.6986
2	110.3232

Now we can plot the means and build a generative additive model using a cubic regression spline bases with 5 knots (represented by the blue line the figure below).



The figure above further confirms our believe that the PM2.5 concentration reaches the lowest during the afternoon at around 3 pm. The figure suggests that the PM2.5 concentration is at its peak during around midnight, gradually decays until afternoon (around 3 pm), and starts increasing again until the next midnight. In addition, the shaded area of the gam regression line which represents the 95% confidence level interval is fairly small.

Final generative additive model

Finally, we will build a model for PM2.5 concentration with all interested variables as the predictors, i.e., hour of the day, wind speed, dew point, and temperature. Among those four predictors, we will add a cubic spline bases with 5 knots to the hour of the day, as it is not naturally linear (hour 0 and 23 are neighbors).

First, we can examine the parametric coefficients of the predictors without spline bases:

Table 8: Coefficients of wind speed, dew point, and temperature

Wind speed	Dew point	Temperature
-0.2590401	5.474609	-6.595348

The coeffiecients of wind speed, dew point, and temperature perfectly aligns with our previous results:

1. for every 1 m/s increase in wind speed, we expect a 0.26 μg/m^3 decrease in PM2.5 concentration;
2. for every 1 °C increase in dew point (more moisture in the air), we expect a 5.47 mug/m^3 increase in PM2.5 concentration;
3. for every 1 °C increase in temperature, we expect a 6.60 μg/m^3 decrease in PM2.5 concentration.

Finally, the R^2 value of the model is 0.249, meaning that it explains 24.9% of the variances. In addition, the p-values for the coefficients are all less than 2×10^{-16}, meaning that there is strong evidence against the hypothesis that the predictors are meaningless.

Decision tree on PM2.5 concentration level

Recall that earlier in the *Data cleaning and wrangling* section, we have created categorical varibles for PM2.5 concentrations, i.e., "good", "moderate", "unhealthy", "very unhealthy", and "hazardous". Now we will build decision trees that utilizes meteological factors to predict this variable. The reason that we choose this method is that decision trees are generally easier to interpret and can be used as a good travel guideline for everyone.

First, we will split our dataset into training (70%) and testing (30%). This will allow us to test the performance of our decision tree without the need to worry about overfit.

After splitting the dataset, we can now build and prune the decision tree. The response variable of decision tree is the PM2.5 concentration level, and the predictors are dew point, temperature, wind speed, and time of the day ("morning", "afternoon", or "evening").

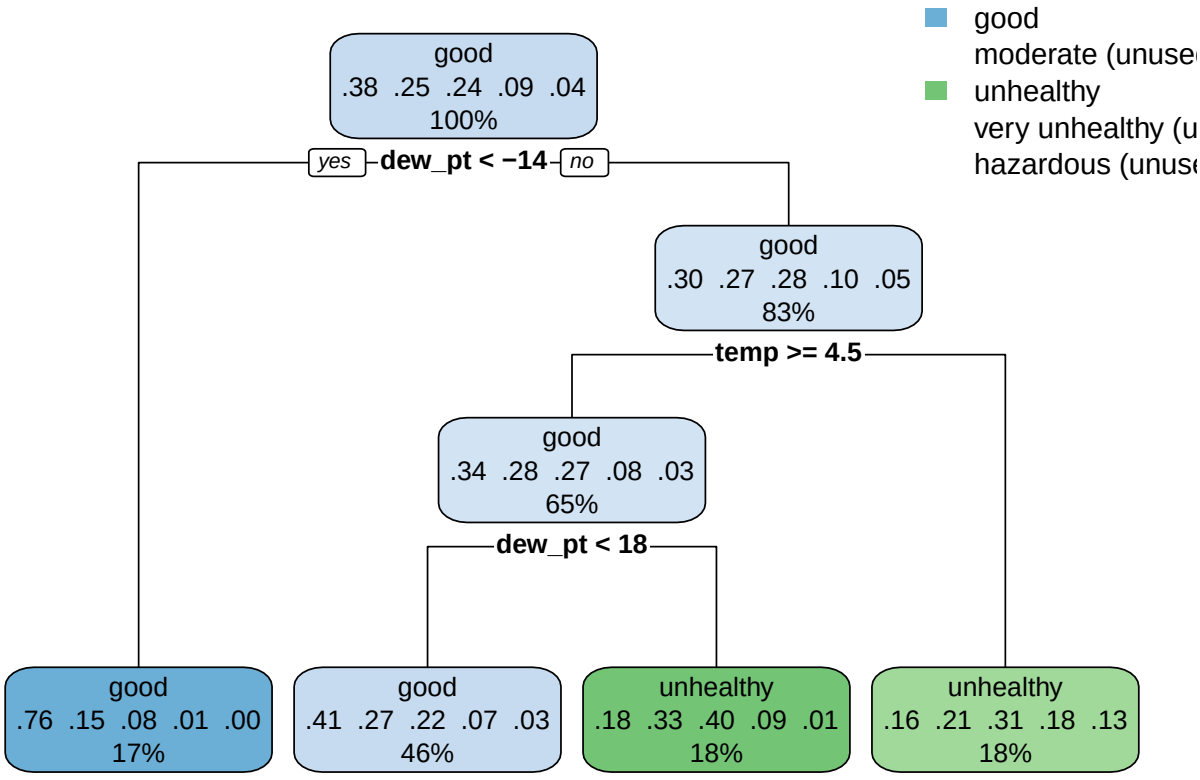


Figure 1: PM2.5 concentration level decision tree

Since the decision tree is simple enough, we do not need to prune it. Since the tree was built using only the training dataset, we can now use it to predict air quality (PM2.5 concentration levels) on the test dataset and evaluate its performance. The proportion of correct predictions in the test dataset is 44%, which is adequate as guessing would yield a 20% accuracy.

From the decision tree plot above, we can see that the decision boundaries aligns perfectly with our previous results from the regression model. To be specific, the decision boundaries show that lower dew point and higher temperature indeed leads to lower PM2.5 concentrations.

Random forest on PM2.5 concentration level

Now we can improve our model by building a random forest consisting many simple decision trees to make more accurate predictions on PM2.5 concentration levels. After building the random forest, we find that accuracy on the test dataset is now 50%.

Conclusion and Summary

Throughout the data exploration, we have used statistical summaries, data visualizations, and generative additive models to answer the question about the associations between PM2.5 concentrations and other meteorological features. To conclude, the data analysis suggests that PM2.5 concentration is the lowest during windy afternoons with dry air (low dew point) and high temperature.

In addition to the above information, we have also successfully built a random forest prediction model with 50% accuracy on the test dataset. Given that our categorization of air quality is rather fine with 5 categories, this model accomplishes the goal of helping people predict PM2.5 concentration levels using time and forecastable meteological information.

References

Liang, X., Zou, T., Guo, B., Li, S., Zhang, H., Zhang, S., Huang, H. and Chen, S. X. (2015). Assessing Beijing's PM2.5 pollution: severity, weather impact, APEC and winter heating. Proceedings of the Royal Society A, 471, 20150257.

What is PM2.5 and why you should care - bliss air. (n.d.). Retrieved February 28, 2022, from https://blissair.com/what-is-pm-2-5.htm