hw3

Andi Liao

January 27, 2019

The major source of air polluants in China is sulfur dioxide emitted from factories, because the energy structure is currently dominated by coal. The central government has released a series of policies to control the concentration and emissions of sulfur dioxide. Although AQI is not the best indicator of sulfur dioxide polluation, areas highly polluted by sulfur dioxide are usually suffer from other air polluants. The paired violin plots illustrate that the air quality index decreases a bit from 2008 to 2011, but then increases greatly from 2011 to 2014. Using the average sulfur dioxide level as cutoff, provinces with higher sulfur dioxide levels have a higher increment in terms of AQI from 2011 to 2014.

It seems that the distribution of polluants are heterogeneous. The dot plot shows the average air quality index grouped by provinces using the data from 2018, 2011 and 2014. The plot indicates that most southern provinces have lower average air quality index value, except for Tibet, which lies in the west plateau region. This finding can be explained by the fact that southern provinces have much denser population, more prosperous economy, warmer climate and higher coverage of vegetation. However, the air pollution is a nationalwide issue, and according to the different circumstances in each province, local governments are supposed to take different actions to help fight against air pollution. We will look at the geo-spatial persperctive more carefully when the time-series data is introduced.

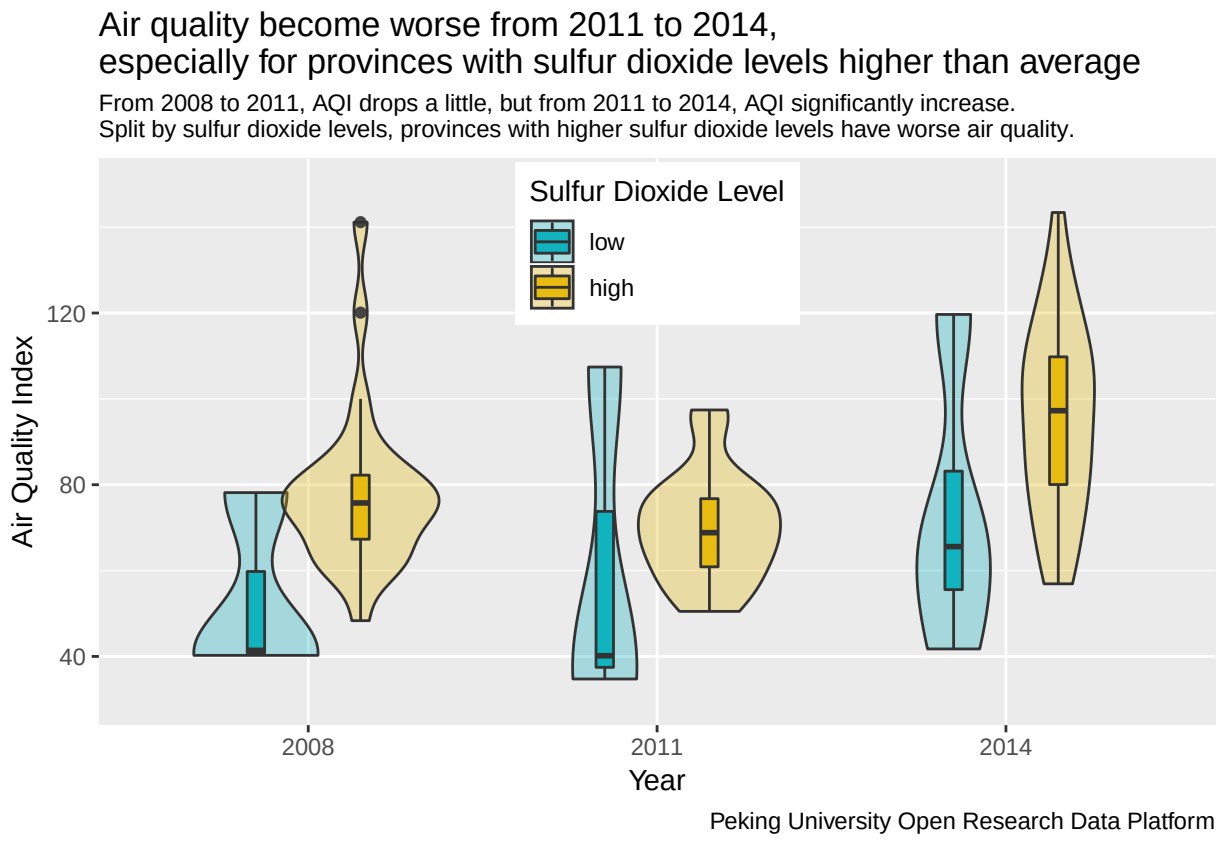
To validate the hypothesis that the severity of air pollution is related with economy and environment, the bubble chart connecting popluation and forest coverage is presented. When population is smaller than average, the relationship between popluation and forest is positive, and it becomes negative when popualtion is larger than average. Interestingly, provinces with worse air quality all have relatively less forest coverage, but their population vary from low to high. The natrual environment factor plays a key role in air quality here.

Let's concentrate on the capital city, Beijing for the next two graphs. Beijing lies in the northern part of China, and it is the most populated city in the country. It is also well-known for sandstrom and haze, but the local government has made efforts to reduce the air pollution for the past decades. From the daily average AQI of Beijing in 2018, it can be seen that most of the time the air is fine (below 100, i.e., no need to wear masks). However, there always some terrible days within a single month in the winter and spring, when the local government often take actions to control pollution by published regulations.

Beijing government started a campagin called "changing from coal to gas" in the winter of 2017. The air quality during winter was largerly improved, compared to previous years. However, Beijing still has a long way to go. Comparing the daily average AQI, PM 2.5 and PM 10 in 2018, PM 2.5 and PM 10 are the major contributors to air pollution, thus they have similiar changing tendencies. There was also a breakout of PM 10 in the spring, but the frequency was much lower than the last decade. The cleaning of time series data is in the process, and the comparasion across years will be used to manifest the efforts made by the local government.

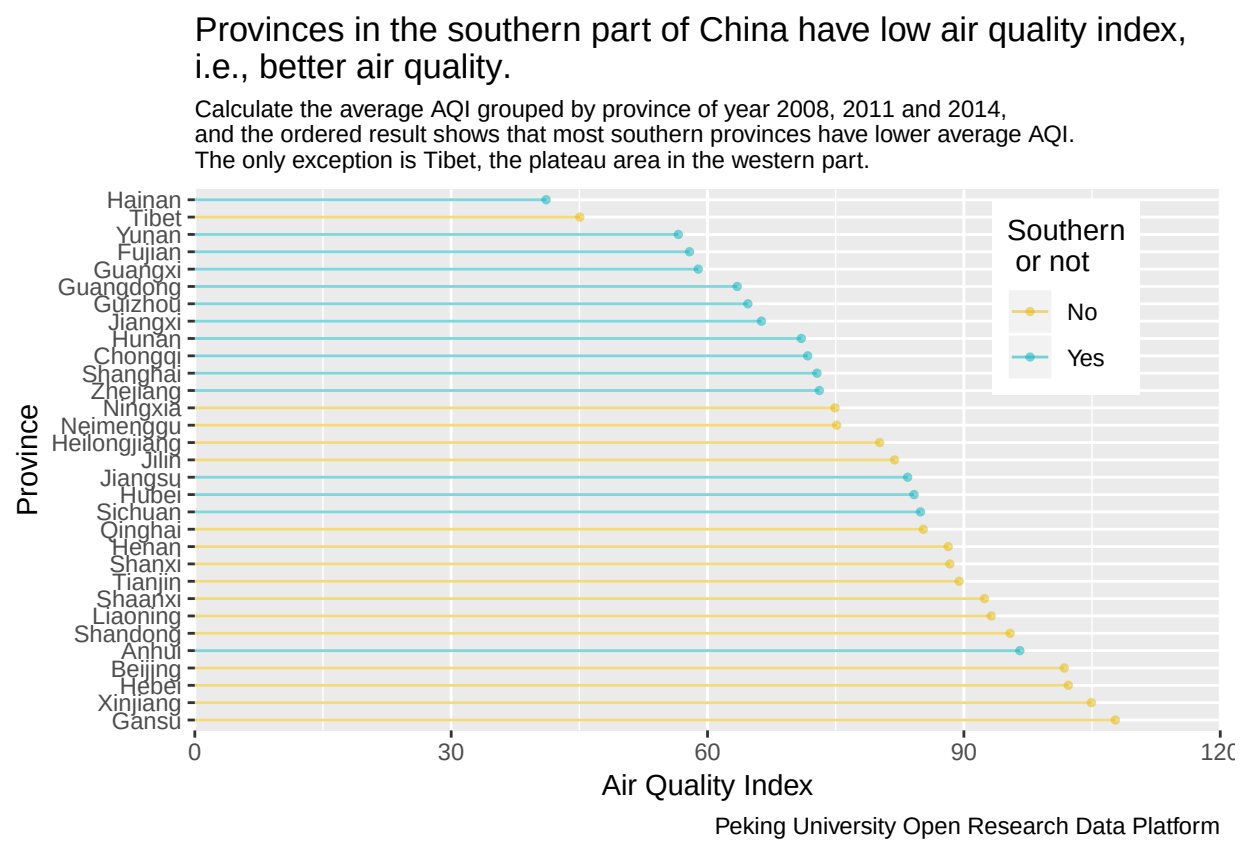
```
data_long %>%
  select(-("metrics")) %>%
  filter(measurement == "aqi" | measurement == "sulfur") %>%
  spread(measurement, value) %>%
  mutate(category = cut(sulfur, breaks=c(0, 70, Inf),labels = c("low", "high"))) %>%
  group_by(category) %>%

  ggplot(aes(x = year, y = aqi, fill = category)) +
  geom_violin(position = position_dodge(width = 0.6), width = 0.9, alpha = 0.3) +
  geom_boxplot(position = position_dodge(width = 0.6), width = 0.1, alpha = 0.9) +
  scale_fill_manual(values = c("#00AFBB", "#E7B800", "#FC4E07")) +
```

```
  ylim(30, 150) +
  theme(plot.subtitle = element_text(size = 9), legend.position = c(0.5, 0.85)) +
  labs(
    title = "Air quality become worse from 2011 to 2014,
especially for provinces with sulfur dioxide levels higher than average",
    subtitle = "From 2008 to 2011, AQI drops a little, but from 2011 to 2014, AQI significantly increas
Split by sulfur dioxide levels, provinces with higher sulfur dioxide levels have worse air quality.",
    caption = "Peking University Open Research Data Platform",
    x = "Year",
    y = "Air Quality Index",
    fill = "Sulfur Dioxide Level"
  )
```



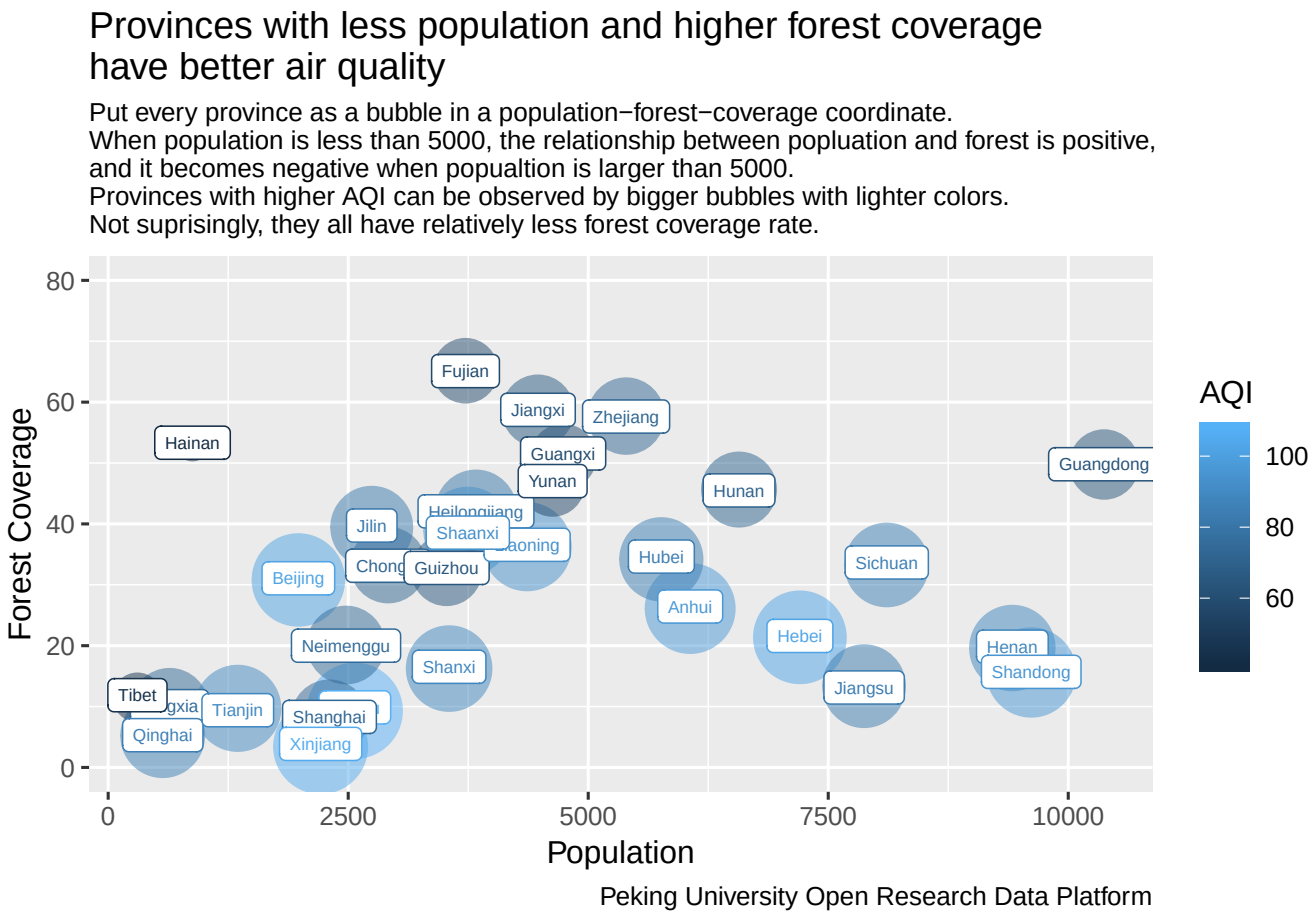

```
p <-
data_long %>%
  select(-("metrics")) %>%
  filter(measurement == "aqi") %>%
  spread(measurement, value) %>%
  group_by(region) %>%
  summarise(aqi = mean(aqi)) %>%
  rownames_to_column("province") %>%
  arrange(aqi) %>%
  mutate(region = factor(region, levels = .$region)) %>%
  mutate(is_south = south_prov)

p %>%
```

```
  ggplot(aes(x = aqi, y = region)) +
  geom_segment(aes(x = 0, xend = aqi, y = region, yend = region,
                   color = factor(is_south)), alpha = 0.5) +
  geom_point(aes(color = factor(is_south)), size = 1, alpha = 0.5) +
  scale_x_continuous(expand = c(0, 0), limits = c(0, 120)) +
  scale_y_discrete(limits = rev(levels(p$region))) +
  scale_color_manual(labels=c("No","Yes"),
                     values = c("#E7B800", "#00AFBB")) +
  theme(plot.subtitle = element_text(size = 9), legend.position = c(0.85, 0.8)) +
  labs(title = "Provinces in the southern part of China have low air quality index,
i.e., better air quality.",
       subtitle = "Calculate the average AQI grouped by province of year 2008, 2011 and 2014,
and the ordered result shows that most southern provinces have lower average AQI.
The only exception is Tibet, the plateau area in the western part.",
       caption = "Peking University Open Research Data Platform",
       x = "Air Quality Index",
       y = "Province",
       color = "Southern\n or not")
```




```
p <-
  data_long %>%
  select(-("metrics")) %>%
  filter(measurement == "aqi" | measurement == "forest" | measurement == "population") %>%
  spread(measurement, value) %>%
  group_by(region) %>%
  summarise(aqi = mean(aqi), forest = mean(forest), population = mean(population))
```

```
p %>%
  ggplot(aes(x = population, y = forest, size = aqi, color = aqi)) +
  geom_jitter(width = 0.5, height = 0.5, alpha = 0.5) +
  scale_size_continuous(range = c(5, 15), guide = FALSE) +
  ylim(0, 80) +
  scale_fill_gradient2(midpoint = 3) +
  geom_label(label = p$region, nudge_x = 0.25, nudge_y = 0.1, size = 2) +
  theme(plot.subtitle = element_text(size = 9)) +
  labs(
    title = "Provinces with less population and higher forest coverage
have better air quality",
    subtitle = "Put every province as a bubble in a population-forest-coverage coordinate.
When population is less than 5000, the relationship between popluation and forest is positive,
and it becomes negative when popualtion is larger than 5000.
Provinces with higher AQI can be observed by bigger bubbles with lighter colors.
Not suprisingly, they all have relatively less forest coverage rate.",
    caption = "Peking University Open Research Data Platform",
    x = "Population",
    y = "Forest Coverage",
    color = "AQI"
  )
```



```
old %>%
  filter(metrics == "AQI") %>%
  group_by(date) %>%
  summarise(AQI = mean(value)) %>%
  ggplot(aes(x = date, y = AQI)) +
  geom_bar(stat = "identity", fill = alpha("blue", 0.3)) +
  coord_polar(start = 0) +
  scale_x_discrete(breaks = seq(20180101, 20181231, by = 60)) +
  ylim(-20, 200) +
  labs(title = "The air quality of 2018 in Beijing is great from July to September,
but mostly terrible in the rest of the year.",
       subtitle = "The x axis should be similar to clock.
This is the daily average air quality index value of Beijing in 2018.
The air quality is great from July to September, the left-bottom part.
However, for the rest part, there always bad days within a single month.",
       caption = "http://beijingair.sinaapp.com/",
       x = "Date",
       y = "AQI")
```


```
old$date = factor(old$date)
old %>%
  spread(metrics, value) %>%
  group_by(date) %>%
  summarise(PM10 = mean(PM10), PM2.5 = mean(PM2.5), AQI = mean(AQI)) %>%
```

```
  arrange(date) %>%
  gather(key = "metrics", value = "value", PM10:AQI) %>%
  ggplot(aes(x = factor(date), y = value, color = metrics))  +
  geom_line(group = 1) +
  facet_grid(metrics ~ .) +
  scale_x_discrete(breaks = seq(20180101, 20181231, by = 60)) +
  labs(title = "The tendencies of AQI, PM 2.5 and PM 10 of Beijing in 2018 are similar,
though there was a breakout of PM10 in the spring.",
       subtitle = "This is the daily average air quality index value, PM 2.5 AND PM 10 concentration of
PM 2.5 and PM 10 are indeed highly correlated with air quality.
There was a breakout of PM 10 around Mar 28th, due to the sandstrom.",
       caption = "http://beijingair.sinaapp.com/",
       x = "Date",
       y = "Concentration")
```