Quantitative Analysis Report for Air Quality 2015 Clinton, Gladstone QLD.

github.com/nixsiow/CSA_113

Yun Kai Siow, 9598138

12 June, 2016

Contents

1. Aim

Question

$PM_{2.5}$ is particulate matter with an equivalent aerodynamic diameter of 2.5 micrometres or less, and generally describe as fine particles. Fine particle like $PM_{2.5}$ can be drawn deep into lungs, bypassing nose and mouth. Because of this, continuously over exposed it can cause bad effects and diseases to human health such as aggravation of asthma or other respiratory system damages. Elderly and children are among those who face higher risk. However, PM~2.5 are very light and tiny, that make it very easy to disturb by wind or carry within wind current.

This study is to investigate what influence do meteorological measurements such as wind speed and wind direction have on the quality of air, particularly concentrations of $PM_{2.5}$?

---

2. Methods

The scientific conceptual model

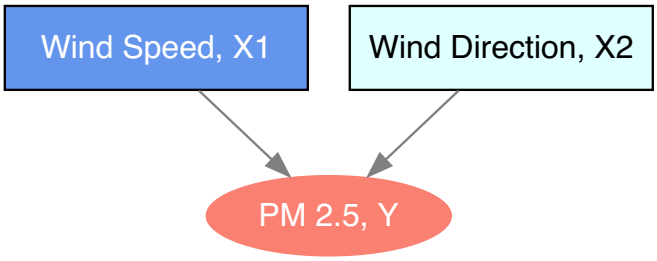

Figure 1: Visual conceptual model of how $PM_{2.5}$ concentration in the air varies according to wind speed and wind direction

As with other meteorological conditions, the explanatory variables, wind speed and wind direction are believe to have play an important role on direct or indirect correlation with the dispersion of air pollutant (e.g. $PM_{2.5}$) concentration in the air ((**???**), (**???**)) (Dawson et al., 2007; Elminir, 2005).

For instance, if there is a forest fire happening at the south west of our current location, a gust of south western wind with the right speed will certainly bring the pollutant, therefore increase the pollutant concentration in the air. Vice versa, wind from other direction with certain speed could also carry away and disperse pollutants in the air.

There is no specified functional form from a scientific law to describe the influence of wind speed and wind direction affect the concentration of $PM_{2.5}$ in the air, so linear terms is used in this model.

The quatitative model

$$\begin{aligned} \log PM_{2.5i} &= \sum_{j=1}^{J} \beta_j \cdot I\left(WD_i = j\right) + \sum_{k=1}^{K} \gamma_k \cdot WS_i \cdot I\left(WD_i = k\right) + \epsilon_i \\ \epsilon_i &\sim N(0, \sigma^2) \end{aligned} \tag{1}$$

Variables & symbols on equation (1):

- $\log PM_{2.5i}$: ith observation of $\log PM_{2.5}$ (logarithm transformed)
- WS_i: Wind speed value for observation i
- WD_i: Wind direction value for observation i
- β_j: Partial effect of wind direction (WD_i) on $\log PM_{2.5}$

- γ_k: Partial effect of interaction term of wind direction (WD_i) and wind speed (WS_i) on log $PM_{2.5}$
- J & K: Total number of wind direction, 8 (e.g. N, NE, E, SE, S, SW, W, NW)
- $I(\cdot)$: an indicator variable that tell us whether or not the statement inside (that Wind Direction has a particular value) is true.

Formulate a hypothesis:

$$H_0 : \beta_j, \gamma_k = 0$$
$$H_1 : \beta_j, \gamma_k \neq 0$$

Significant value $\alpha = 0.05$

---

3. Data

Preparation

```
# CSV file read from downloaded source from current working
# directory.
air.quality.clinton.raw <- read.csv(file = "data/clinton-aq-2015.csv",
    as.is = T, head = T)

# Or

# Read/download directly from data custodian (Queensland
# Government open data portal).
url <- "http://www.ehp.qld.gov.au/data-sets/air-quality/clinton-aq-2015.csv"
air.quality.clinton.raw <- read.csv(file = url, as.is = T, head = T)
```

Data preparation before analysis

```
air.quality.clinton.raw$Date <- dmy_hm(paste(air.quality.clinton.raw$Date,
    air.quality.clinton.raw$Time))
# Lubridate to add few extra colume: month, day_of_week
library(lubridate)
air.quality.clinton.raw <- mutate(air.quality.clinton.raw, month = month(Date,
    label = T), day_of_week = wday(Date, label = T))

# define data of interest & rearrange the seq & save to new
```

```r
# df
data.of.interest <- c("Date", "Time", "month", "day_of_week",
    "PM2.5..ug.m.3.", "Wind.Speed..m.s.", "Wind.Direction..degTN.")
air.quality.clinton <- subset(air.quality.clinton.raw, select = data.of.interest)

# Rename variables name
names(air.quality.clinton) <- c("date", "time", "month", "day_of_week",
    "pm2.5", "ws", "wd")

# Assign breakboint for cutting and labelling 0 and 360 are
# for NORTH
breaks = c(0, seq(22.5, 337.5, by = 45), 360)

# cut function dplyr to divides the range of feeded data into
# intervals and codes the values in 'Direction' such as NE, E
# ...  according to which interval they fall. Turn the
# continuous variable to categorial variable
library(dplyr)
wd.label <- cut(air.quality.clinton$wd, breaks = breaks, dig.lab = 4,
    labels = c("N", "NE", "E", "SE", "S", "SW", "W", "NW", "N"),
    include.lowest = TRUE)

# Create new categorical variable wd.label Logarithm
# transforms PM2.5 to log.pm2.5 due to data skewness Remove
# na value. Remove negative pm value
air.quality.clinton <- mutate(air.quality.clinton, wd.label = wd.label) %>%
    mutate(log.pm2.5 = log(pm2.5)) %>% na.omit %>% filter(!is.nan(log.pm2.5) &
    !is.infinite(log.pm2.5))

# Check the levels of wd.label
levels(air.quality.clinton$wd.label)
# regroup both 'N' level into only one level, should be only
# 8 instead of 9
levels(air.quality.clinton$wd.label) <- c("N", "NE", "E", "SE",
    "S", "SW", "W", "NW", "N")
```

Dataset

Look at the first few rows of the data to see what is contained within.

```
head(air.quality.clinton)
```

Table 1: First 6 row of the final dataset.

date	time	month	day_of_week	pm2.5	ws	wd	wd.label	log.pm2.5
2015-01-01 01:00:00	01:00	Jan	Thurs	3.4	2.6	58	NE	1.2237754
2015-01-01 02:00:00	02:00	Jan	Thurs	2.1	3.0	63	NE	0.7419373
2015-01-01 04:00:00	04:00	Jan	Thurs	1.2	1.5	82	E	0.1823216
2015-01-01 05:00:00	05:00	Jan	Thurs	6.0	1.0	128	SE	1.7917595
2015-01-01 06:00:00	06:00	Jan	Thurs	5.0	1.6	120	SE	1.6094379
2015-01-01 07:00:00	07:00	Jan	Thurs	4.7	2.5	96	E	1.5475625

Data dictionary

Data dictionary - variables

Table 2: Data dictionary listed with abbreviations, descriptions, units, permissible range of each variables.

Abbreviation	Variable	Description	Units	Permissible range
ws	Wind speed	Measured by ultrasonic sensor with 10 metres above ground level.	$ms^{-}1$	0.1, 12.6
wd.label	Wind direction in 8 catagory	Measured by ultrasonic sensor with 10 metres above ground level.	-	N, NE, E, SE, S, SW, W, NW

Abbreviation	Variable	Description	Units	Permissible range
log.pm2.5	Log transformed $PM_{2.5}$	Particulate matter with an equivalent aerodynamic diameter of 2.5 micrometres or less.	$\mu g/m^3$	-2.3025851, 4.5920849

The final data set comprises time series of wind speed and direction; and $PM_{2.5}$ readings. All updated hourly over the period from 1st January to 31st December 2015, recorded at Clinton, Gladstone Queensland (Latitude: -23.8701; Longitude: 151.2216).

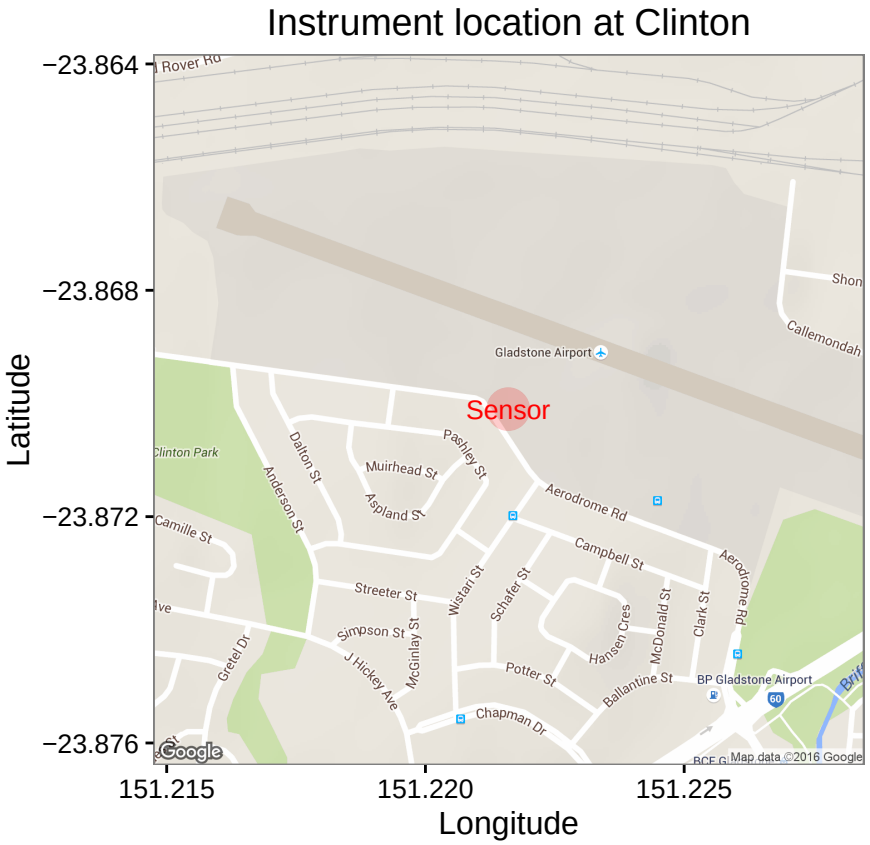

Figure 2: Location of the physical sennsing instrument at Clinton, Gladstone QLD.

Metadata

The dataset is released under a Creative Commons Attribution 3.0 Australia (CC BY) licence.

.

Experimental design and standards

1. **Wind:** The wind speed, X_1 and wind direction, X_2 are measured by ultrasonic sensor with 10 metres above ground level, compliant to Meteorological monitoring for ambient air quality monitoring applications (AS/NZS 3580.14:2011). Wind direction sensor is aligned to magnetic north and the output value of reported wind direction is referenced to true north by application of a magnetic declination correction of +10 degrees.

- **Measurement units:**
  - Wind speed, metres per second ($ms^{-}1$),
  - Wind direction, $degTN$

2. **$PM_{2.5}$:** Particles as $PM_{2.5}$ means particulate matter with an equivalent aerodynamic diameter of 2.5 micrometres or less. The suspended particular matter - $PM_{2.5}$ concentrations are measured by Dichotomous Tapered Element Oscillating Balance (TEOM) Model 1405-DF fitted with Filter Dynamics Measurement System (FDMS) operated in accordance with Method 9.13, Australian Standards Methods for Pollutant Monitoring (AS/NZS 3580.9.13). The FDMS system compensates for the loss of semi-volatile components from the collected particulate matter. Reported concentrations are uncorrected instrument output values and calculated from running 1-hour average concentrations updated at six minute intervals. Negative hourly $PM_{2.5}$ concentrations down to -5$\mu g/m^3$ resulting from instrument noise at low particle concentrations are reported.

- **Measurement units:** micrograms per cubic metre ($\mu g/m^3$)

There is no specified functional form from a scientific law to describe the influence of wind speed and wind direction affect the concentration of $PM_{2.5}$ in the air, so linear terms is used in this model.

---

4. Analysis

Exploratory data analysis

```
summary(air.quality.clinton[,c(6,8,9)])
```

Table 3: Numerical summaries of explanatory variables and outcome variable

ws	wd.label	log.pm2.5
Min. : 0.100	SE :1878	Min. :-2.3026
1st Qu.: 1.700	E :1656	1st Qu.: 0.8755
Median : 2.800	S :1383	Median : 1.3863
Mean : 3.209	SW :1176	Mean : 1.2696
3rd Qu.: 4.400	NE :1098	3rd Qu.: 1.7918
Max. :12.600	N : 408	Max. : 4.5921
NA	(Other): 290	NA

Note that raw $PM_{2.5}$ data is logarithm transformed to get rid of some negative readings and correcting overall skew results.

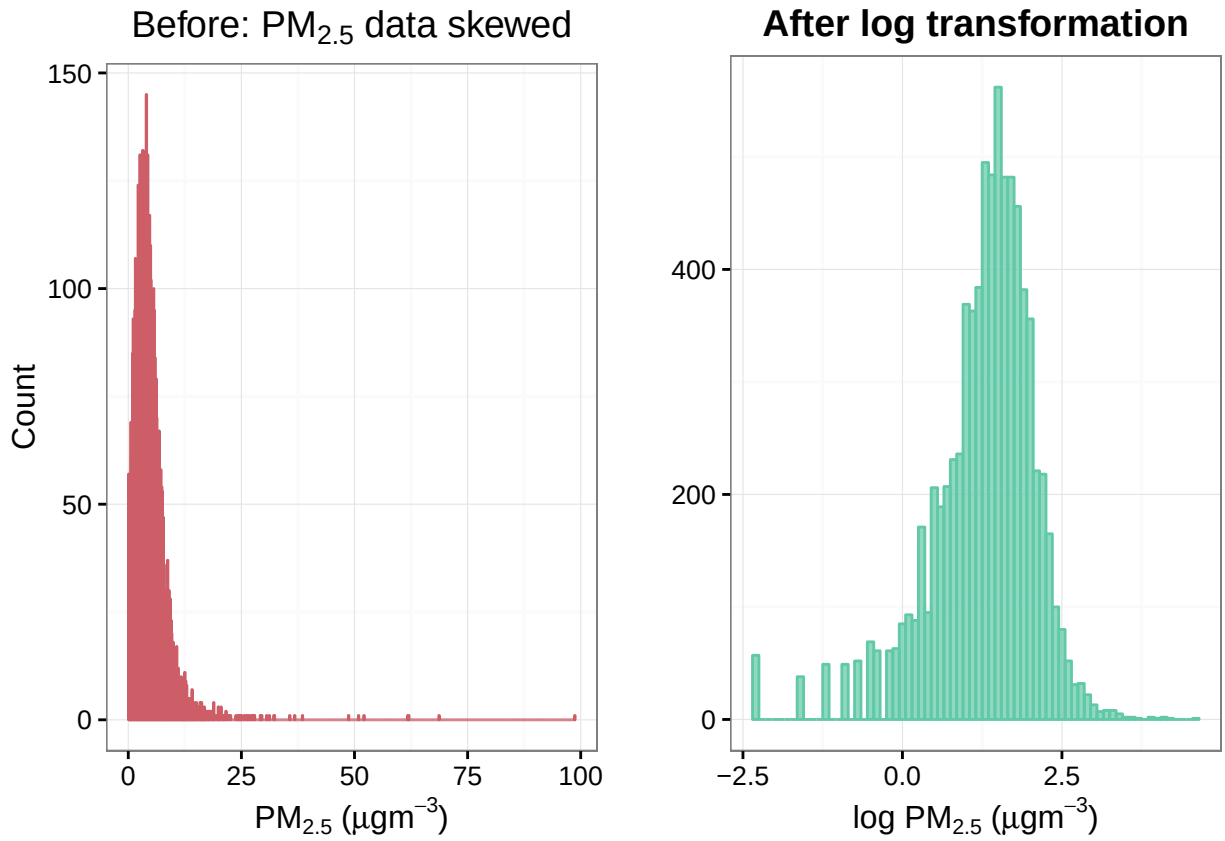

Figure 3: Histogram showing the variation in outcome variable, both PM2.5 and log PM2.5

How many observations fall on each wind category? What the most common wind direction are?

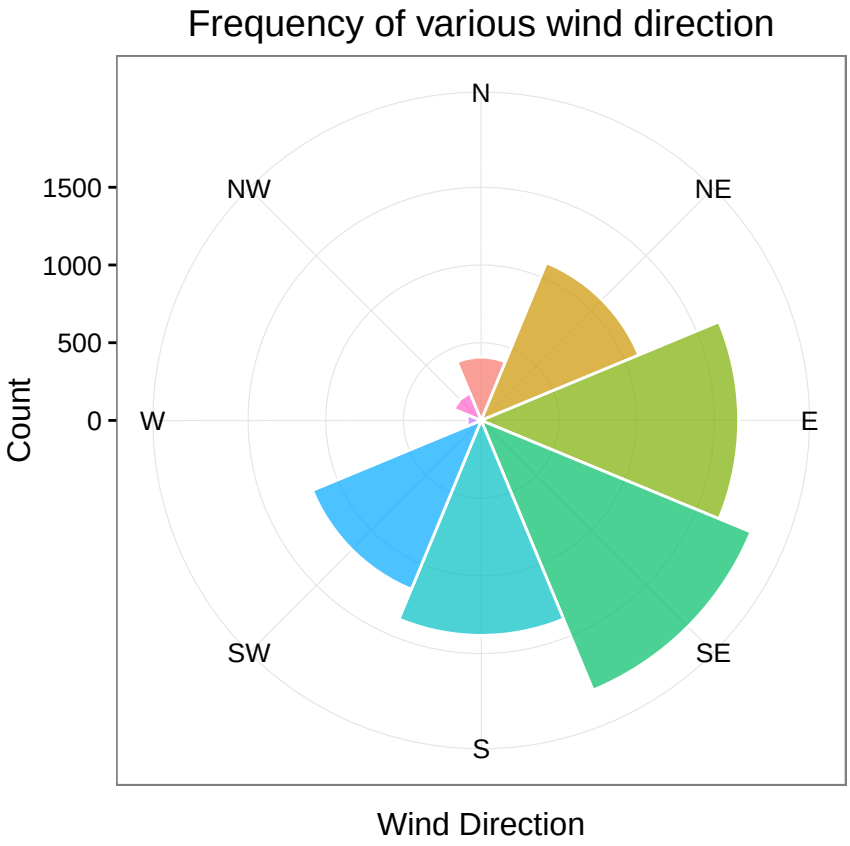

Figure 4: Graph showing the variation in one explanatory variable, Wind Direction

Figure 4 shows that most commonly wind came from East and South Eastern, on the other hand, there are very less wind come from West direction. Wind blowing from different direction can cause dispersal of air pollutants or carry within pollutants from emission source.

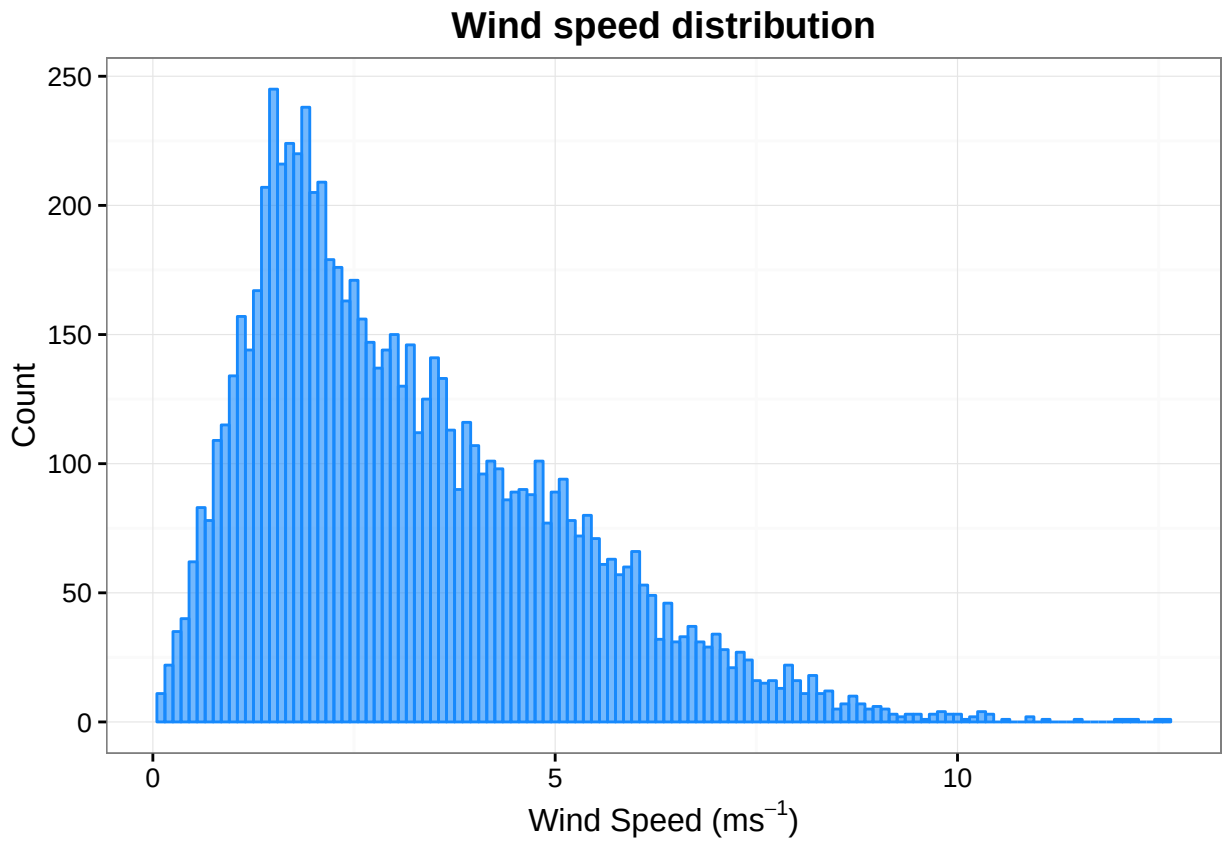

Figure 5: Histogram showing the variation in another explanatory variable, Wind Speed

Figure 5 shows that the wind speed distribution in the observation data. The most common wind speed is around 2 meter per seconds, light winds.

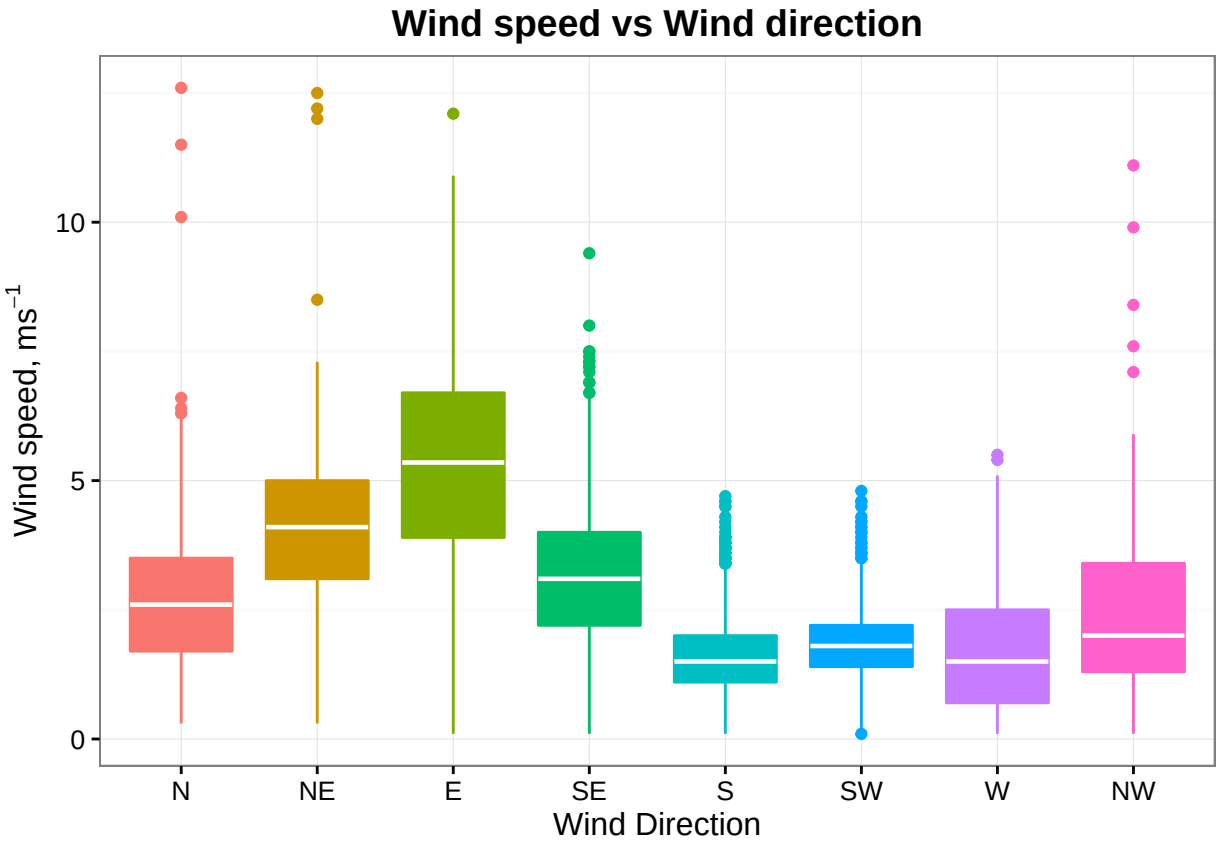

Figure 6: Bar plot showing the correlation of both explanatory variable, Wind Direction and Wind Speed.

In figure 6, clearly the median show that wind from Eastern and North Eastern blow stronger compared to others.

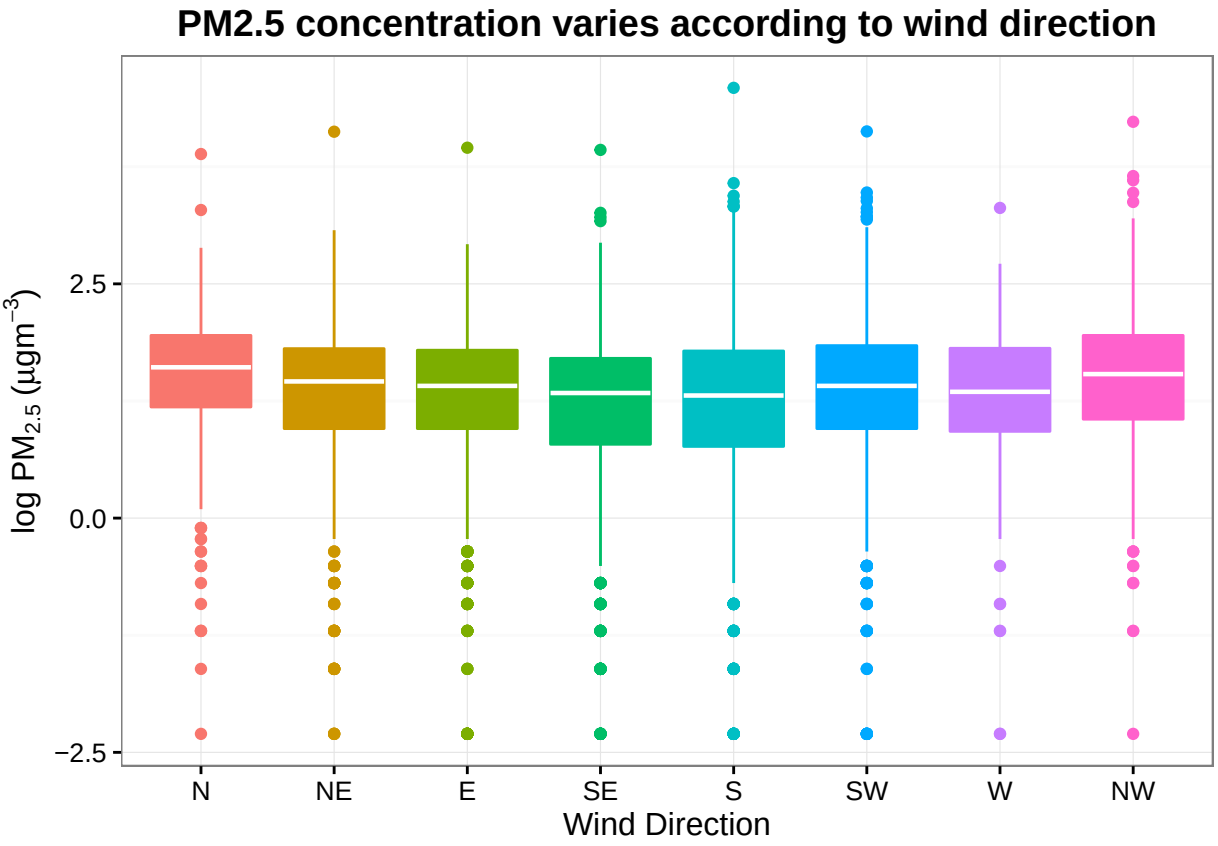

Figure 7: blank

The median $PM_{2.5}$ from North is higher than the others which might indicate that there is more air polution when the wind is blowing from the north than the other directions.

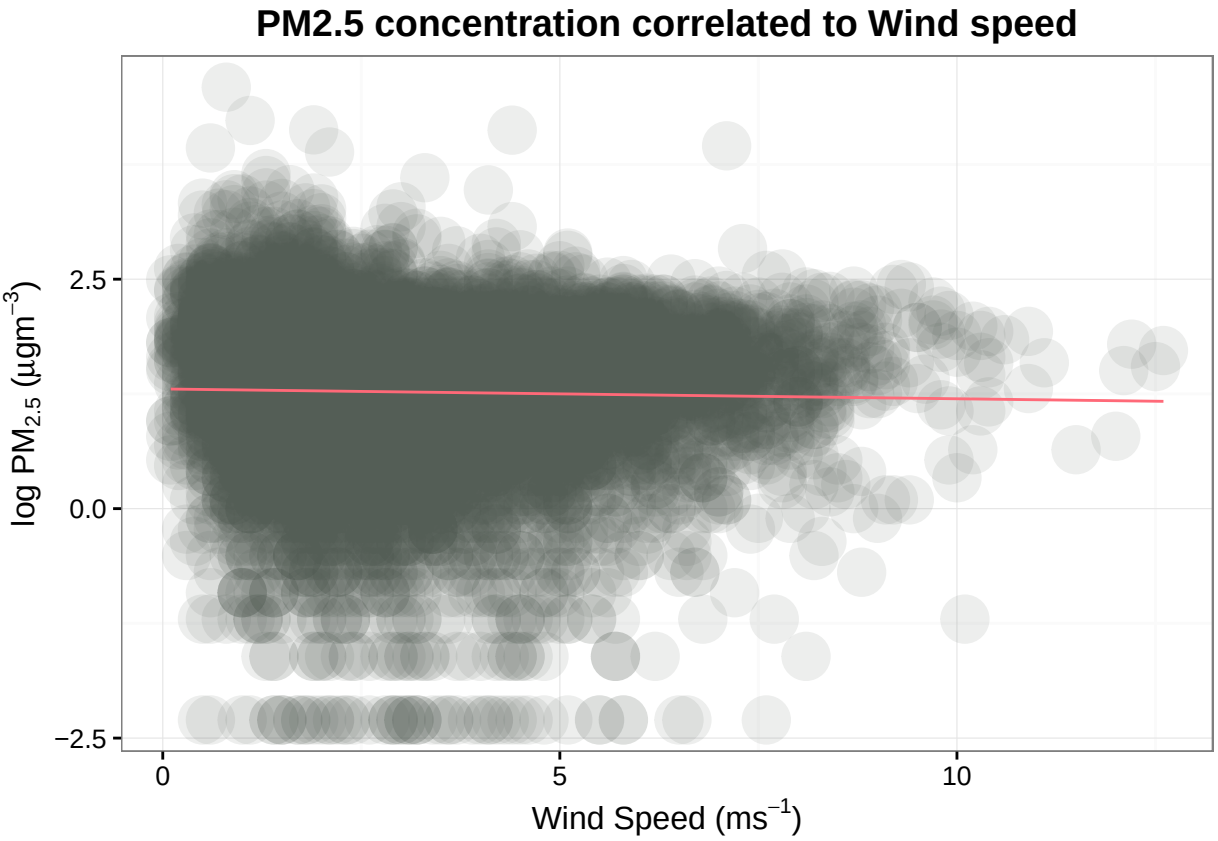

Figure 8: blank

Figure 8 shows that plotting single variable, wind speed alone has very low correlation with PM2.5 data, which then suggest there are others factors in the system might explain more the variability, for example the wind direction.

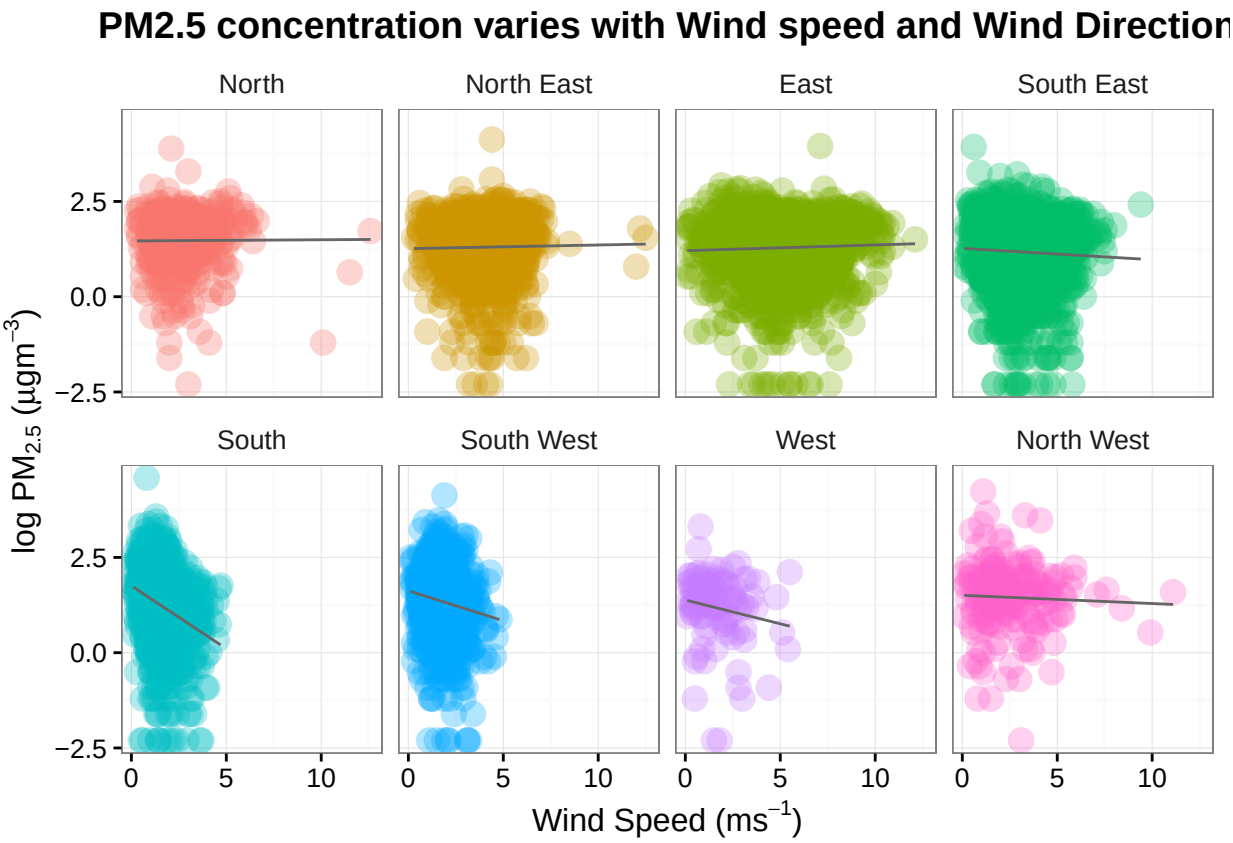

Figure 9: blank

Figure 9 showing how $PM_{2.5}$ varies with both of Wind Speed and 8 differnt Wind Directions. Wind from South and South West seem to have negative correlation with PM~2.5.

Quantitative analysis

Fit 6 models as below:

- Model 1, log.pm2.5 ~ ws
- Model 2, log.pm2.5 ~ wd.label
- Model 3, log.pm2.5 ~ ws+wd.label
- Model 4, log.pm2.5 ~ ws:wd.label
- Model 5, log.pm2.5 ~ wd.label + ws:wd.label
- Model 6, log.pm2.5 ~ ws + wd.label + ws:wd.label

6 models to be fit to explore effect from different paramenters and any extra variability is explaining by adding extra terms or extra parameters.

Model 1
The first model is log $PM_{2.5}$ only predicted by Wind Speed (ws).

```
lm.pm_ws <- lm(data=air.quality.clinton, log.pm2.5 ~ ws)
```

Model 2
The second model is log $PM_{2.5}$ only predicted by Wind Direction (wd.label).

```
lm.pm_wd <- lm(data=air.quality.clinton, log.pm2.5 ~ wd.label - 1)
# Below is same model without setting the intercept to 0. For R2 and ANOVA
lm.pm_wd.intercept <- lm(data=air.quality.clinton, log.pm2.5 ~ wd.label)
```

Model 3
The third model is log $PM_{2.5}$ regressed on both explanatory variables, Wind Speed and Wind Direction.

```
lm.pm_ws_wd <- lm(data=air.quality.clinton, log.pm2.5 ~ ws+wd.label - 1)
# Below is same model without setting the intercept to 0. For R2 and ANOVA
lm.pm_ws_wd.intercept <- lm(data=air.quality.clinton, log.pm2.5 ~ ws+wd.label)
```

Model 4
The fourth model is log $PM_{2.5}$ regressed on only the interaction terms of both explanatory variables, Wind Speed and Wind Direction.

```
lm.pm_wswd <- lm(data=air.quality.clinton, log.pm2.5 ~ ws:wd.label -1)
# Below is same model without setting the intercept to 0. For R2 and ANOVA
lm.pm_wswd.intercept <- lm(data=air.quality.clinton, log.pm2.5 ~ ws:wd.label)
```

Model 5
The fifth model is log $PM_{2.5}$ regressed on Wind Direction and the interaction terms of both explanatory variables, Wind Speed and Wind Direction.

```
lm.pm_wd_wswd <- lm(data = air.quality.clinton, log.pm2.5 ~ wd.label -
    1 + ws:wd.label)
# Below is same model without setting the intercept to 0. For
# R2 and ANOVA
lm.pm_wd_wswd.intercept <- lm(data = air.quality.clinton, log.pm2.5 ~
    wd.label + ws:wd.label)
```

Model 6
The sixth model is log $PM_{2.5}$ regressed on both Wind Direction, Wind Speed and the interaction terms of both explanatory variables, Wind Speed and Wind Direction. This is the most complex model by far.

```
lm.pm_ws_wd_wswd <- lm(data=air.quality.clinton, log.pm2.5 ~ ws*wd.label-1)
# Below is same model without setting the intercept to 0. For R2 and ANOVA
lm.pm_ws_wd_wswd.intercept <- lm(data=air.quality.clinton, log.pm2.5 ~ ws*wd.label)
```

Coefficient of determination, R^2 value of all 6 models

```
library(broom)
# R2 for model 6, repeat for other model
glance(lm.pm_ws_wd_wswd.intercept)$r.squared
```

Check coefficient of determination, R^2 value for each fitted linear models using `glance()` from `broom` library and results tabulated as below:

Table 4: Coefficient of determination for all models

Model	R2s
M1	0.000622744551669358
M2	0.00969867627821189
M3	0.0112528141881653
M4	0.0196769965879333
M5	0.0295696450949947
M6	0.0295696450949948

Table 4 shows that model 5 and model 6 both has the exactly the same R_2 value even though model 6 has extra term. It seems like the extra term (Wind speed) doesn't explain any extra variability in the data.

However, we are going to use F test via the `anova()` function to test whether including the Wind Speed (ws) term improve how much variation in log $PM_{2.5}$ is explained when compared to the model without the Wind Speed (ws) term.

Formulate a hypothesis test about models:

H_0 : There is no increase in variability explained by the more complex model 6, with extra term (wind speed)

Significant value set as $\alpha = 0.05$.

```
# Model 1: log.pm2.5 ~ wd.label + ws:wd.label
# Model 2: log.pm2.5 ~ ws * wd.label
anova(lm.pm_wd_wswd.intercept, lm.pm_ws_wd_wswd.intercept)
```

Table 5: Analysis of Variance Table

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
7873	5180	NA	NA	NA	NA
7873	5180	0	1.819e-12	NA	NA

Interpretation of ANOVA for regression models From table 5, model 5 and model 6 both has 7873 residual degrees of freedom. The p value is not show on the `anova()` result probably due to very tiny differnt between both RSS (regression sum of squared). Therefore, we conclude that model 6 explains slightly more of the varibility (really tiny RSS) but not so much so that the Wind Speed term is necessary. In result, we are unable to reject the null hypothesis, Model 5 the simpler model is the better model choice.

Estimate model parameters

Estimates of the parameters in this model and their 95% confidence intervals

```
tidy(lm.pm_wd_wswd, conf.int = T)
```

Table 6: Confident intervals of estimated parameters of model 5.

term	estimate	std.error	statistic	p.value	conf.low	conf.high
wd.labelN	1.4636466	0.0851682	17.1853617	0.0000000	1.2966943	1.6305989
wd.labelNE	1.2609472	0.0725992	17.3686116	0.0000000	1.1186335	1.4032609
wd.labelE	1.2079854	0.0564363	21.4044169	0.0000000	1.0973553	1.3186154
wd.labelSE	1.2687922	0.0487256	26.0395422	0.0000000	1.1732771	1.3643073
wd.labelS	1.7539877	0.0520973	33.6675442	0.0000000	1.6518632	1.8561122
wd.labelSW	1.6250921	0.0653020	24.8857926	0.0000000	1.4970829	1.7531014
wd.labelW	1.3821692	0.1395471	9.9046780	0.0000000	1.1086198	1.6557186
wd.labelNW	1.5064520	0.1027556	14.6605307	0.0000000	1.3050237	1.7078802
wd.labelN:ws	0.0030367	0.0273998	0.1108304	0.9117536	-0.0506741	0.0567476
wd.labelNE:ws	0.0095950	0.0169719	0.5653487	0.5718527	-0.0236744	0.0428644
wd.labelE:ws	0.0149528	0.0099839	1.4976964	0.1342522	-0.0046182	0.0345239
wd.labelSE:ws	-0.0298001	0.0141242	-2.1098649	0.0349015	-0.0574871	-0.0021130
wd.labelS:ws	-0.3296870	0.0291472	-11.3110970	0.0000000	-0.3868233	-0.2725507
wd.labelSW:ws	-0.1574398	0.0333043	-4.7273139	0.0000023	-0.2227250	-0.0921545
wd.labelW:ws	-0.1247721	0.0667685	-1.8687271	0.0616979	-0.2556560	0.0061119

term	estimate	std.error	statistic	p.value	conf.low	conf.high
wd.labelNW:ws	-0.0216569	0.0340720	-0.6356196	0.5250429	-0.0884471	0.0451334

From the p value of table 7, except `wd.labelN:ws`, `wd.labelNE:ws`, `wd.labelE:ws`, `wd.labelW:ws`, `wd.labelNW:ws`, all the other estimated parameters has shown statistical significant on explaining varibility of the data. This 5 statistical insignificant estimated parameters also shown to have include `0` in their 95% confident intervals.

Assess model fit

Goodness of fit plot

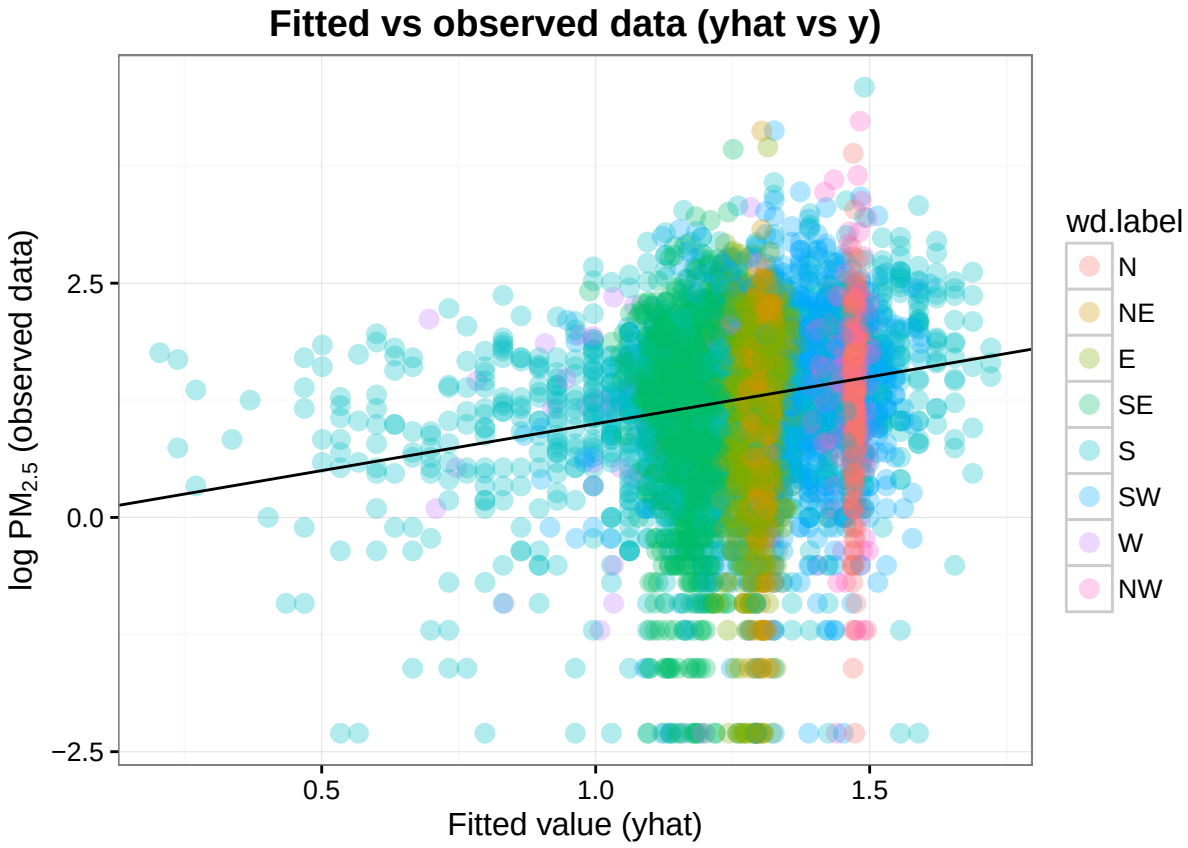

Figure 10: Graph shows how much do our modelled value look like our observed values.

Figure 10 shown how much do our modelled values look like our observed log $PM_{2.5}$ values. They are expected them to fall very close to a straight line if the modelled values is very good explaining variability of the observed data.

At the end of the line, there is wide spread of modelled value along the line. This incompactible suggested that there are still alot of variability haven't been explained by our explanatory variables.

Model checking

```
# Check lm whether the residuals are normally distributed
df.fort.pm_wd_wswd <- fortify(lm.pm_wd_wswd)
head(df.fort.pm_wd_wswd)
```

Table 7: fortify of model 5

log.pm2.5	wd.label	ws	.hat	.sigma	.cooksd	.fitted	.resid	.stdresid
1.2237754	NE	2.6	0.0018025	0.8111534	0.0000007	1.285894	-0.0621189	-0.0766549
0.7419373	NE	3.0	0.0013727	0.8111302	0.0000392	1.289732	-0.5477950	-0.6758351
0.1823216	E	1.5	0.0027784	0.8110674	0.0002916	1.230415	-1.0480931	-1.2939826
1.7917595	SE	1.0	0.0019804	0.8111297	0.0000577	1.238992	0.5527673	0.6821773
1.6094379	SE	1.6	0.0012944	0.8111419	0.0000186	1.221112	0.3883258	0.4790733
1.5475625	E	2.5	0.0017820	0.8111465	0.0000155	1.245367	0.3021951	0.3729058

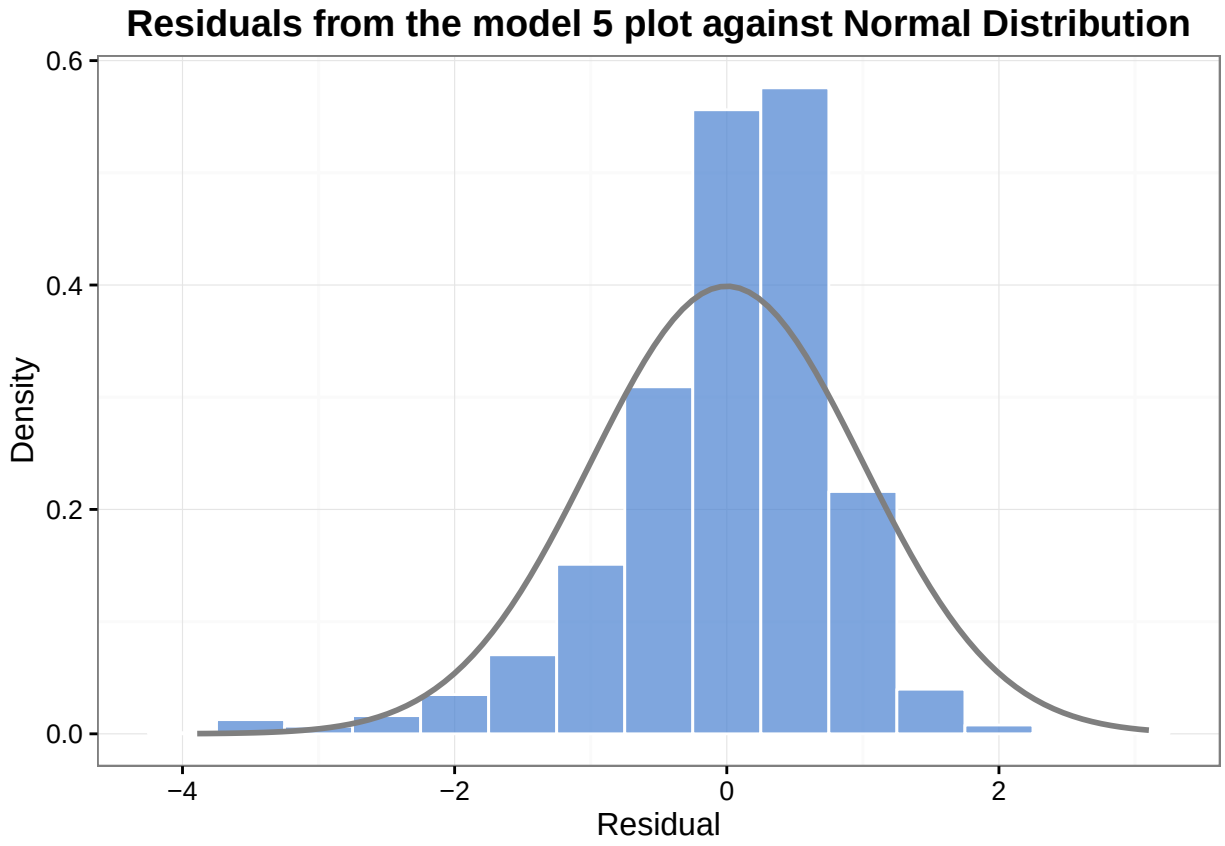

Figure 11: Histogram of the residuals with density on the y axis with assumption of normal distribution of the residuals

The residuals look approximately normally distributed, but look a little skewed and out of shape around the mean.

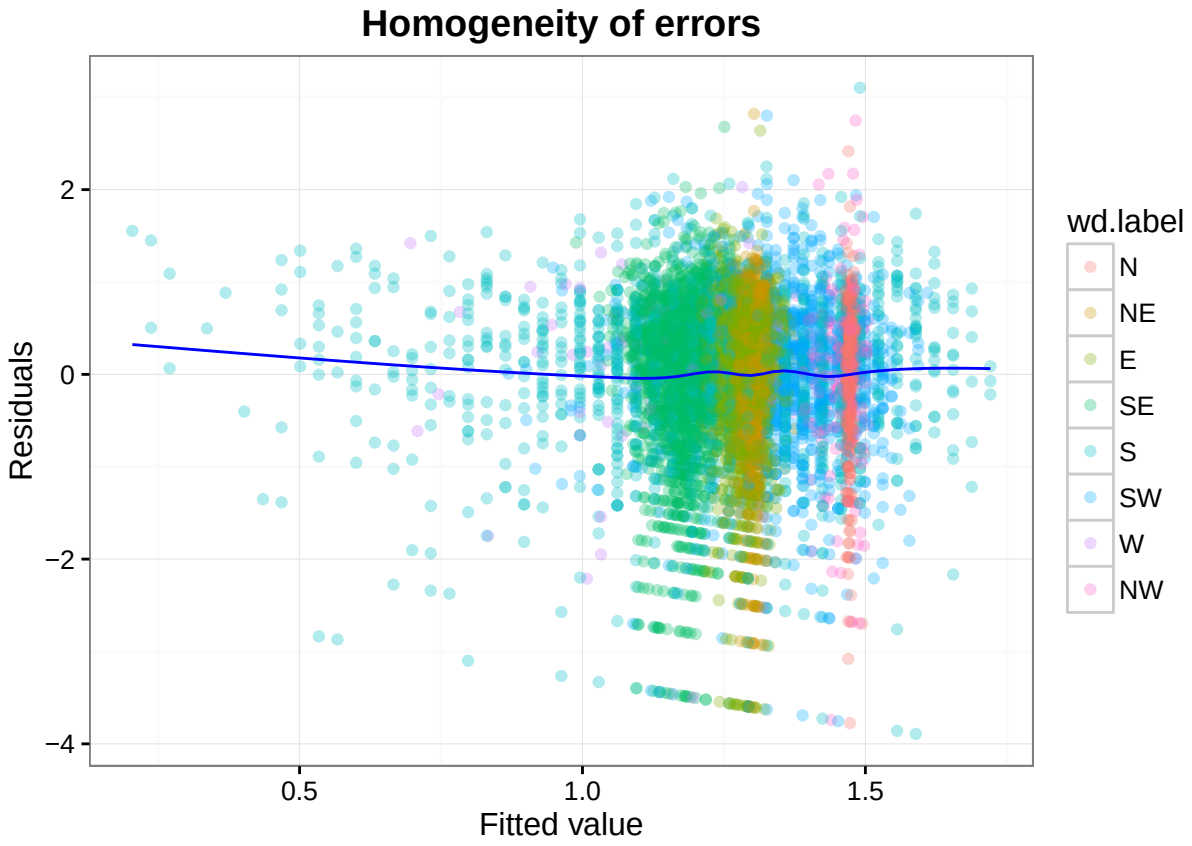

Figure 12: Homogeneity of errors of model 5

The residuals doesn't look like they have a mean of zero and constant variance as we move from left to right along the fitted values axis. This suggested that maybe the residuals are not residuals are normally distributed and homogenous in their variance.

It look like there's many unexplained variation in the residuals. The shape of the plot looks alot like fan out as we move from left to right along the fitted value axis. Most of the residuals concentrate around 1 to 1.5 fitted value.

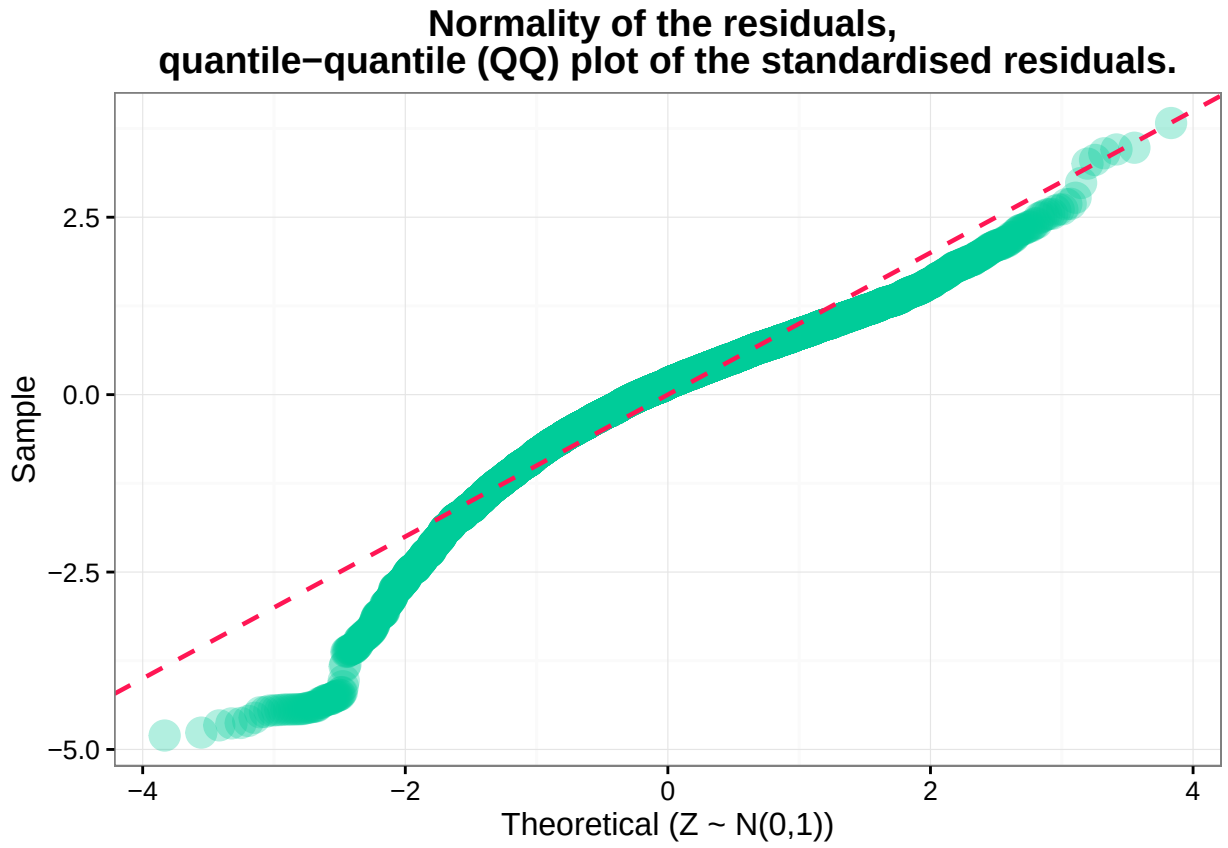

Figure 13: Quantile-quantile (QQ) plot of the standardised residuals.

QQ plot from figure 13 shows that these residuals are not really well Normally distributed due to the skewness from the histograms and the departure from the geom_abline() line sample = theoretical. The sample residuals are less positive than expected as they are way below the line. Perhaps there is some structure to the unexplained variation in the residuals.

---

5. Interpret

Model interpretation

It's very clear that wind speed and wind direction has play a role in explaining log $PM_{2.5}$ based on our fitted model. The estimated parameters are statistical significant under our significal value 0.05.

However, the fitted model explain not very much of the variability of the data based on low R^2 value and the model ND assumption is not hold very well when we look at the residuals.

That is to say, this model while is the best fitted so far, but its not enough to make prediction or inferences on future $PM_{2.5}$ value.

From the result, it suggested that thare is other factors in the system for example temperature might play a role in further predict the $PM_{2.5}$ concentration in the air.

This study answer the original question of whether meteorological measurements such as wind speed and wind direction have on the quality of air, particularly concentrations of $PM_{2.5}$. Further include other measurement might improve the overall explanation of variability.

All analyses were conducted using the statistical software program, R (R Core Team, 2016).

D. Kahle and H. Wickham. (Kahle & Wickham, 2013): Spatial Visualization with ggplot2

References

Kahle, D., & Wickham, H. (2013). Ggmap: Spatial visualization with ggplot2. *The R Journal*, *5*(1), 144–161. Retrieved from http://journal.r-project.org/archive/2013-1/kahle-wickham.pdf

R Core Team. (2016). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/