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

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

Diagram

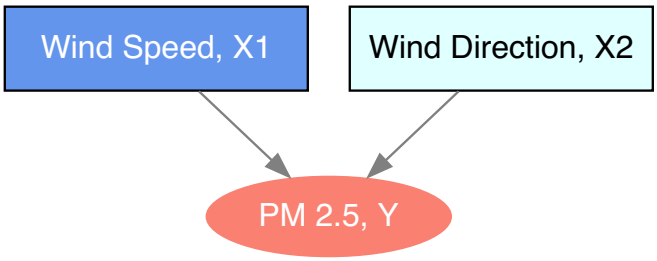

Figure 1: Visual conceptual model of how $PM_{2.5}$ concentration in the air varies according to wind speed and wind direction

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:

$$H_0 : \beta_1 = 0$$
$$H_1 : \beta_1 \neq 0$$

---

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)
```

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

date	time	month	day_of_week	pm2.5	ws	wd	wd.label	log.pm2.5
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
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).

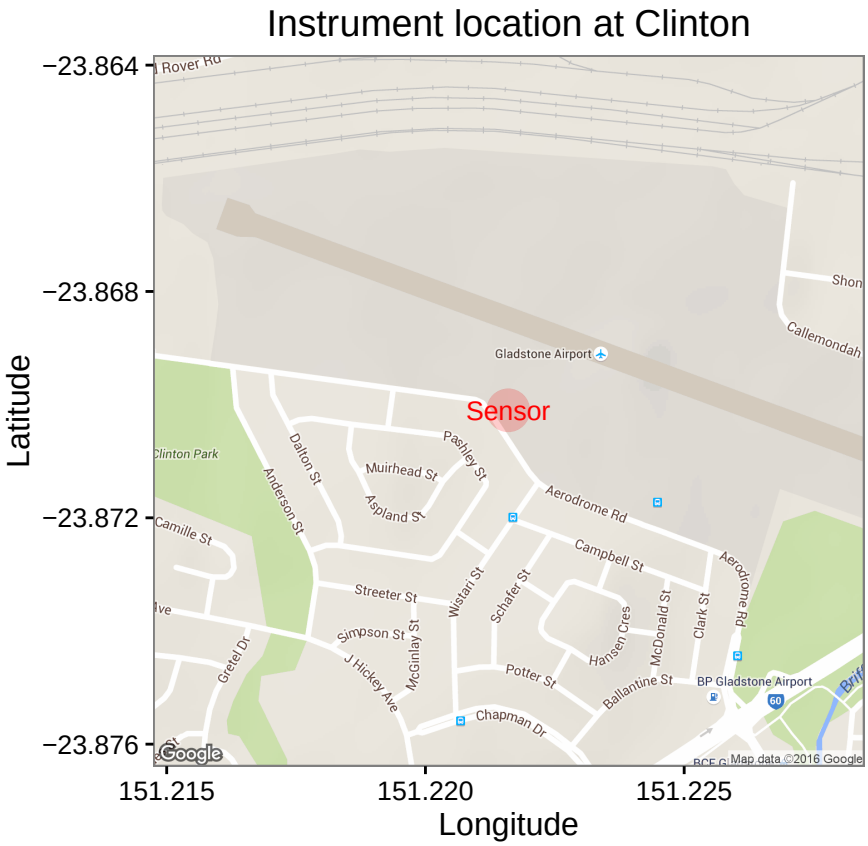

Figure 2: Location of the physical sennsing instrument at Clinton.

Metadata

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

Table 3: summary of explanatory variables and outcome variable

pm2.5	ws	wd.label
Min. : 0.10	Min. : 0.100	SE :1878
1st Qu.: 2.40	1st Qu.: 1.700	E :1656
Median : 4.00	Median : 2.800	S :1383
Mean : 4.68	Mean : 3.209	SW :1176
3rd Qu.: 6.00	3rd Qu.: 4.400	NE :1098
Max. :98.70	Max. :12.600	N : 408
NA	NA	(Other): 290

$PM_{2.5}$ is logarithm transformed to get rid of skew results.

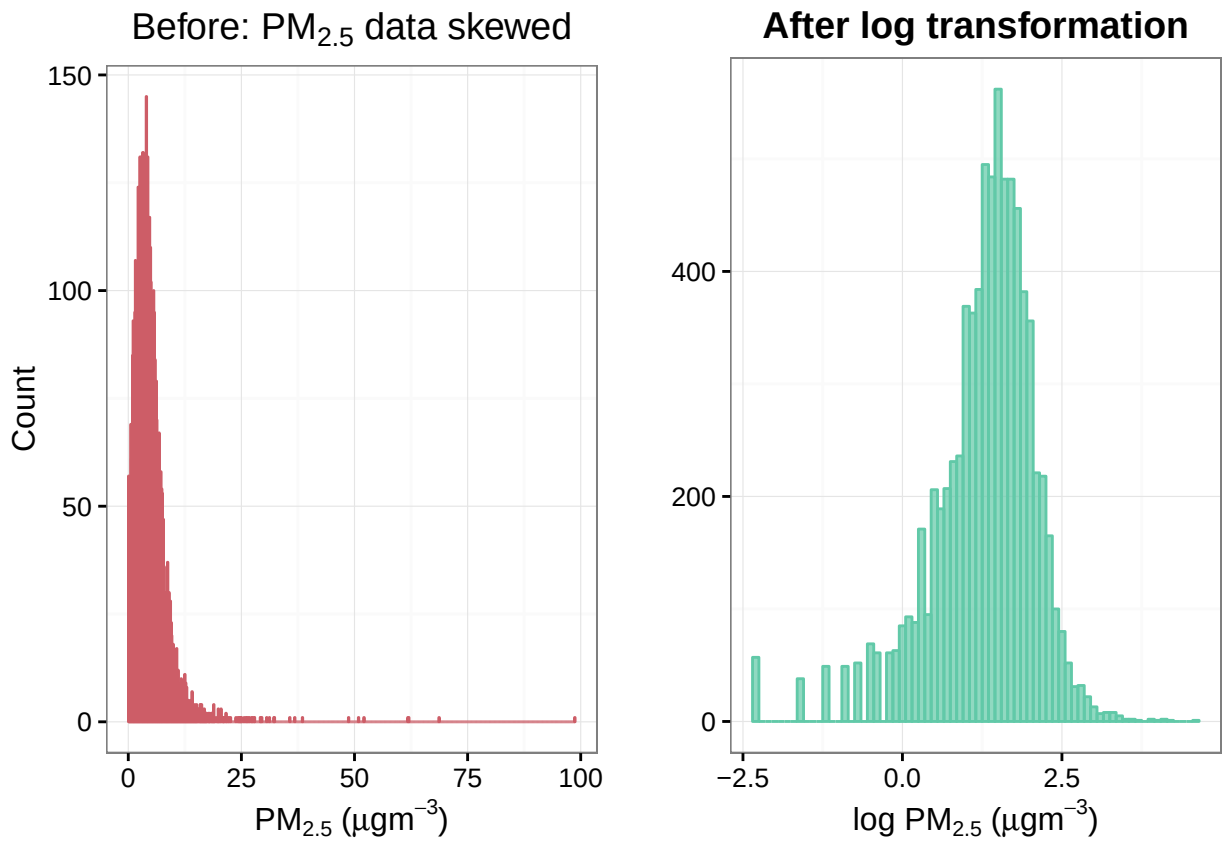

Figure 3: blank

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

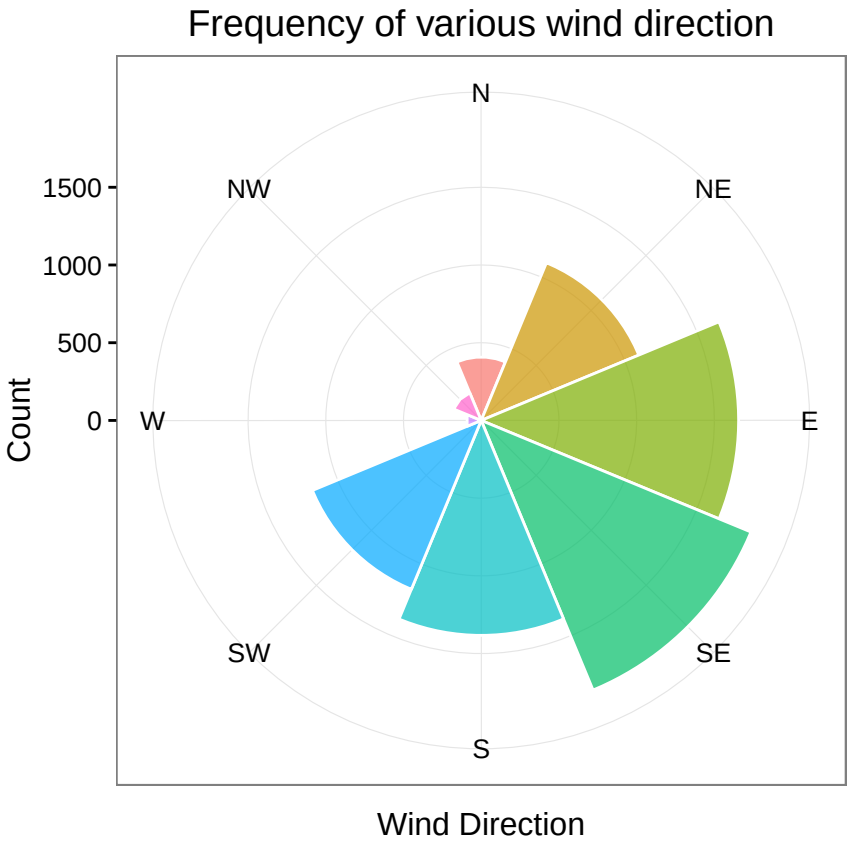

Figure 4: blank

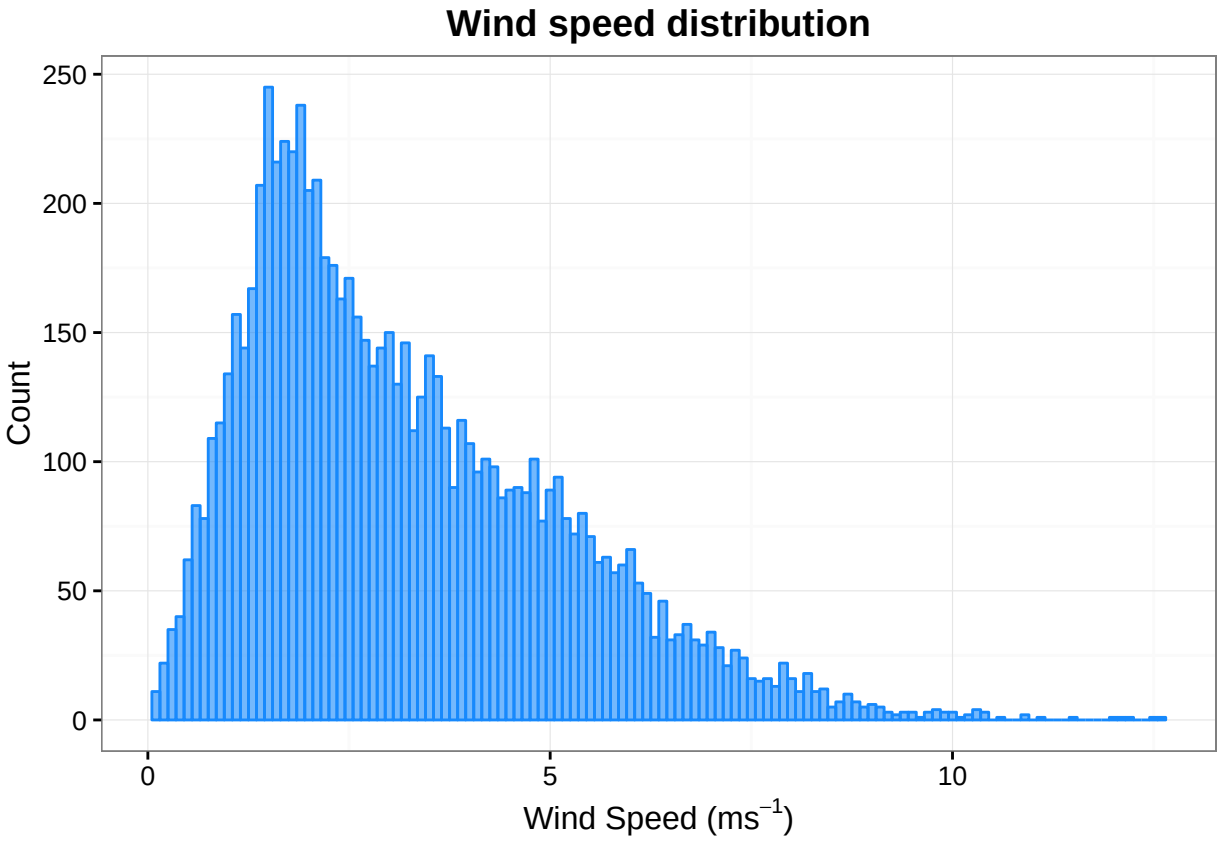

Figure 5: blank

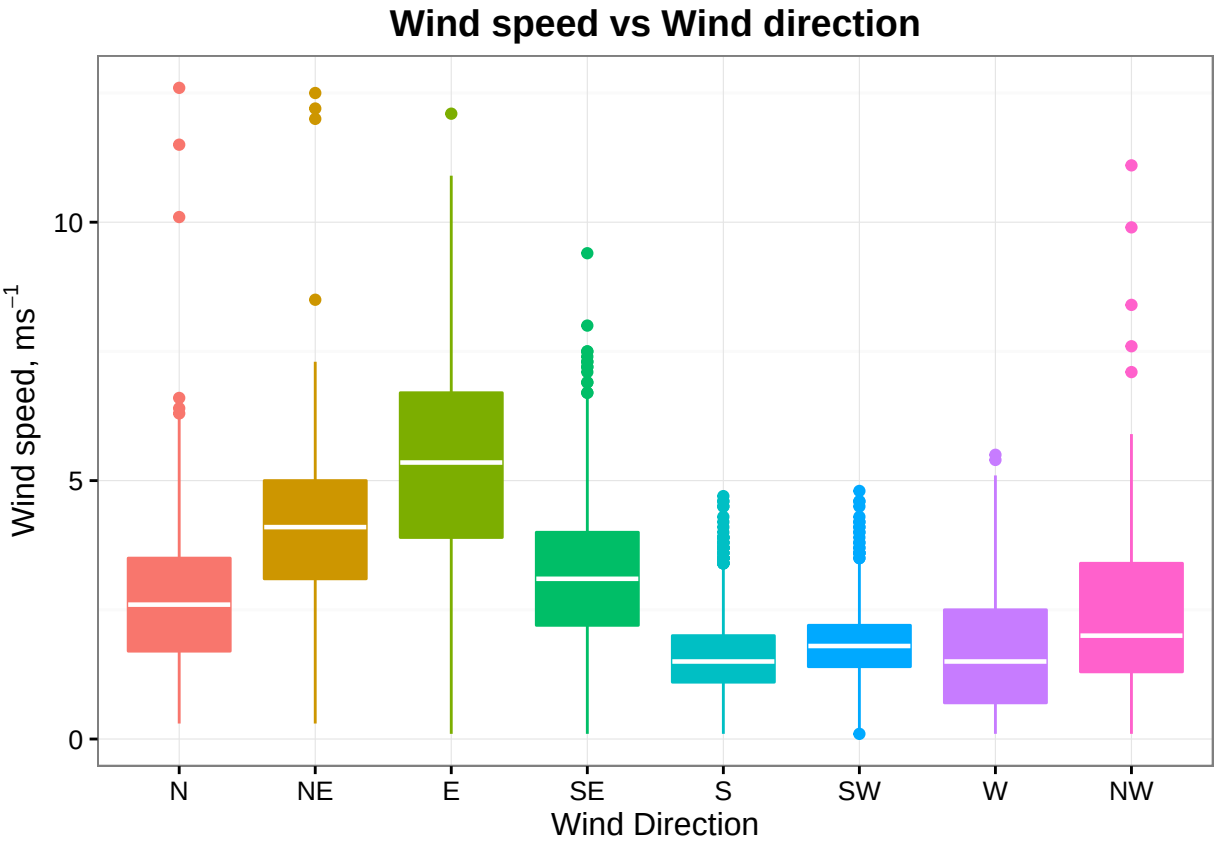

Figure 6: blank

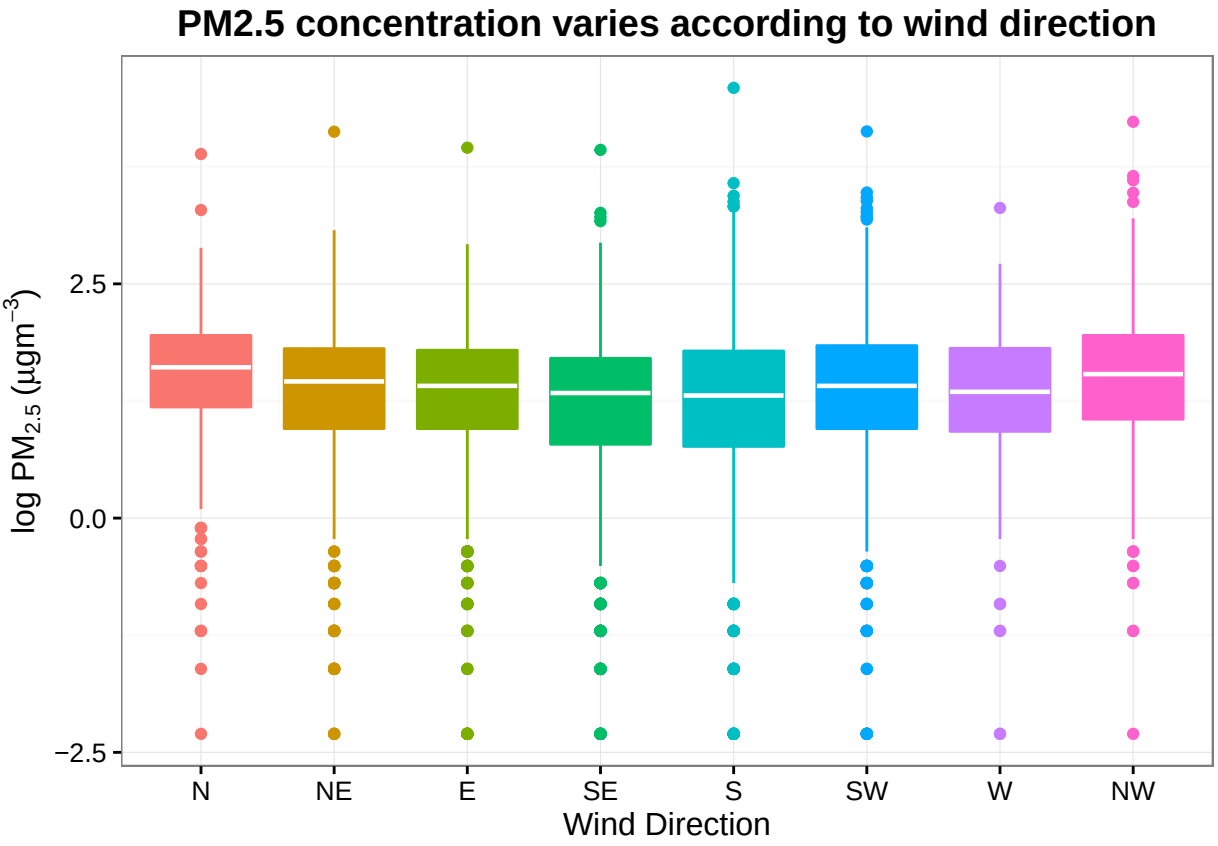

Figure 7: blank

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

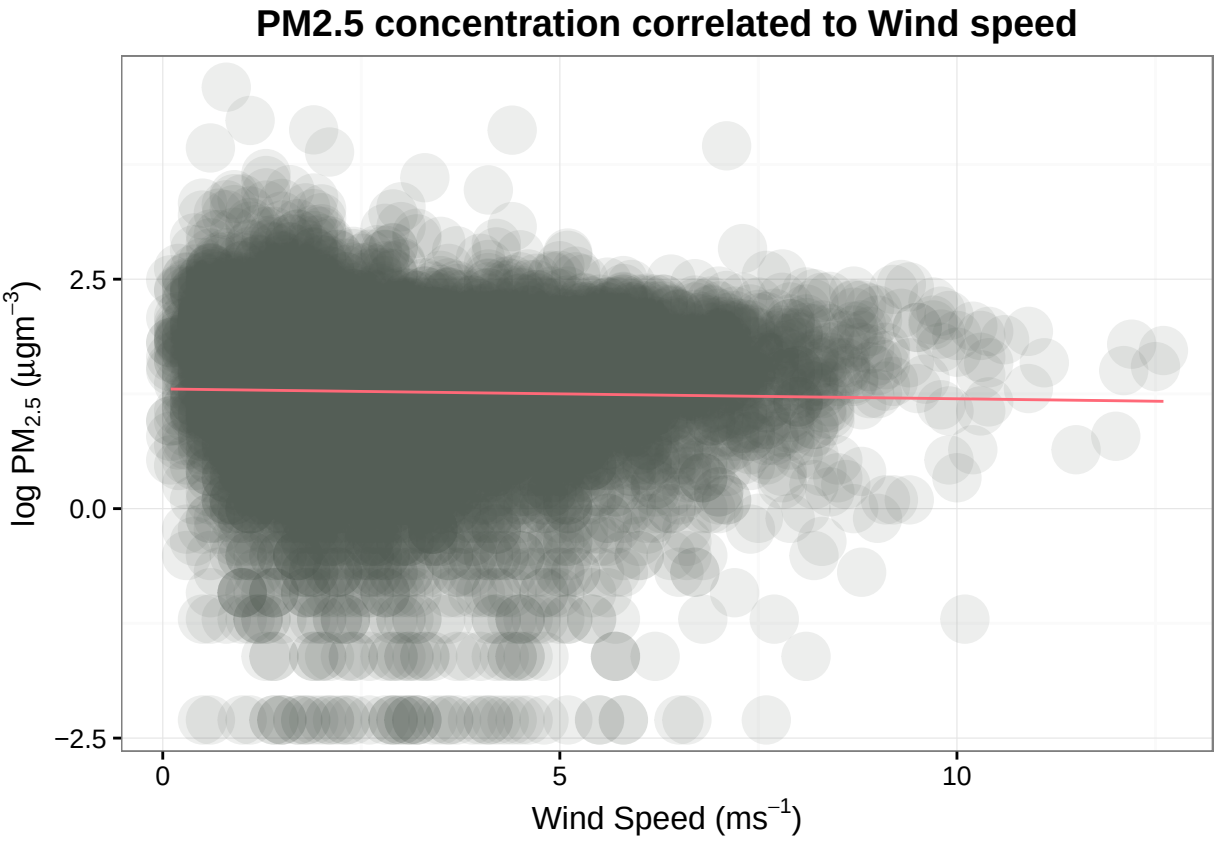

Figure 8: blank

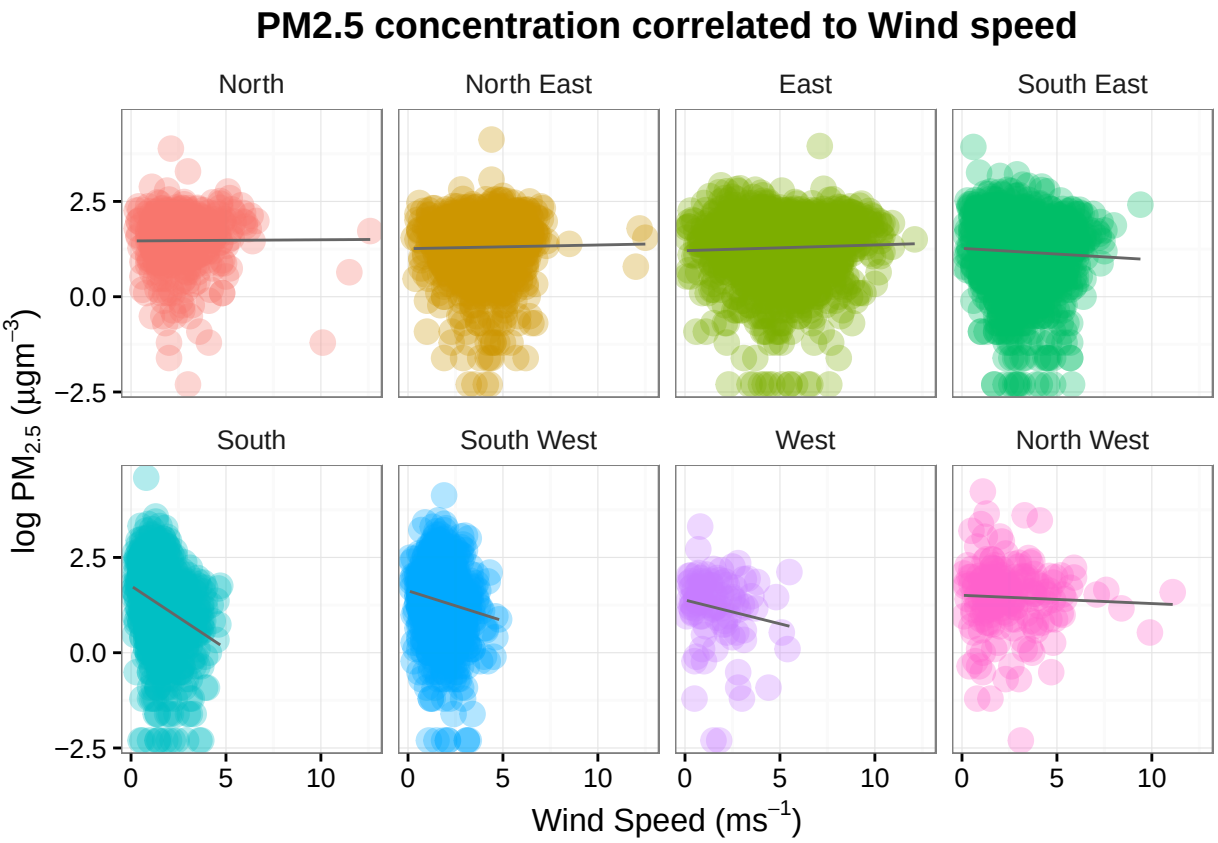

Figure 9: blank

Quantitative analysis

Fit 6 models as follow: 1. log.pm2.5 ~ ws 2. log.pm2.5 ~ wd.label 3. log.pm2.5 ~ ws+wd.label 4. log.pm2.5 ~ ws:wd.label 5. log.pm2.5 ~ wd.label + ws:wd.label 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 parameters.

Model 1

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

Model 2

```
lm.pm_wd <- lm(data=air.quality.clinton, log.pm2.5 ~ wd.label - 1)
lm.pm_wd.intercept <- lm(data=air.quality.clinton, log.pm2.5 ~ wd.label)
```

Model 3

```
lm.pm_ws_wd <- lm(data=air.quality.clinton, log.pm2.5 ~ ws+wd.label - 1)
lm.pm_ws_wd.intercept <- lm(data=air.quality.clinton, log.pm2.5 ~ ws+wd.label)
```

Model 4

```
lm.pm_wswd <- lm(data=air.quality.clinton, log.pm2.5 ~ ws:wd.label -1)
lm.pm_wswd.intercept <- lm(data=air.quality.clinton, log.pm2.5 ~ ws:wd.label)
```

Model 5 Fit a linear model of log $PM_{2.5}$ regressed on Wind direction and interaction with Wind speed

```
lm.pm_wd_wswd <- lm(data=air.quality.clinton, log.pm2.5 ~ wd.label-1 + ws:wd.label)
lm.pm_wd_wswd.intercept <- lm(data=air.quality.clinton, log.pm2.5 ~ wd.label + ws:wd.lab
```

Model 6

```
lm.pm_ws_wd_wswd <- lm(data=air.quality.clinton, log.pm2.5 ~ ws*wd.label-1)
lm.pm_ws_wd_wswd.intercept <- lm(data=air.quality.clinton, log.pm2.5 ~ ws*wd.label)
```

Coefficient of determination, R^2 of all 6 models

Table 4: R2 for all models

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

Formulate a hypothesis test about models: H_0 : There is no increase in variability explained by the more complex model 5, with extra term (wind direction)

Significant value set as $\alpha = 0.05$.

Interpretation of ANOVA for regression models The p value from this hypothesis test therefore represents the probability of obtaining an F statistic at least as big as what was seen if the restricted model explained the same amount of variation as the full model.

Rejecting the null hypothesis leads us to conclude that the inclusion of the extra terms in the full model explains more variation than if we had not included these terms.

```
# model 4 nested inside of model 5
```

```
anova(lm.pm_wswd.intercept, lm.pm_wd_wswd.intercept)
```

```
## Analysis of Variance Table
##
## Model 1: log.pm2.5 ~ ws:wd.label
## Model 2: log.pm2.5 ~ wd.label + ws:wd.label
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1   7880 5232.3
## 2   7873 5179.5  7    52.801 11.465 1.459e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# Model 5 vs. Model 6, model 5 nested in model 6
anova(lm.pm_wd_wswd.intercept, lm.pm_ws_wd_wswd.intercept)
```

```
## Analysis of Variance Table
##
## Model 1: log.pm2.5 ~ wd.label + ws:wd.label
## Model 2: log.pm2.5 ~ ws * wd.label
##   Res.Df    RSS Df Sum of Sq F Pr(>F)
## 1   7873 5179.5
## 2   7873 5179.5  0 1.819e-12
```

That is, does including an interaction term explain any further variation than changes in Wind direction and Wind speed would account for by themselves?

Estimate model parameters

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

```
summary(lm.pm_wd_wswd)
```

95% Confident interval

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

Table 5: Confident intervals of estimated parameters.

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

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

Assess model fit

Coefficient of determination, R^2, for thess models

Table 6: R2 for all models

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

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)
```

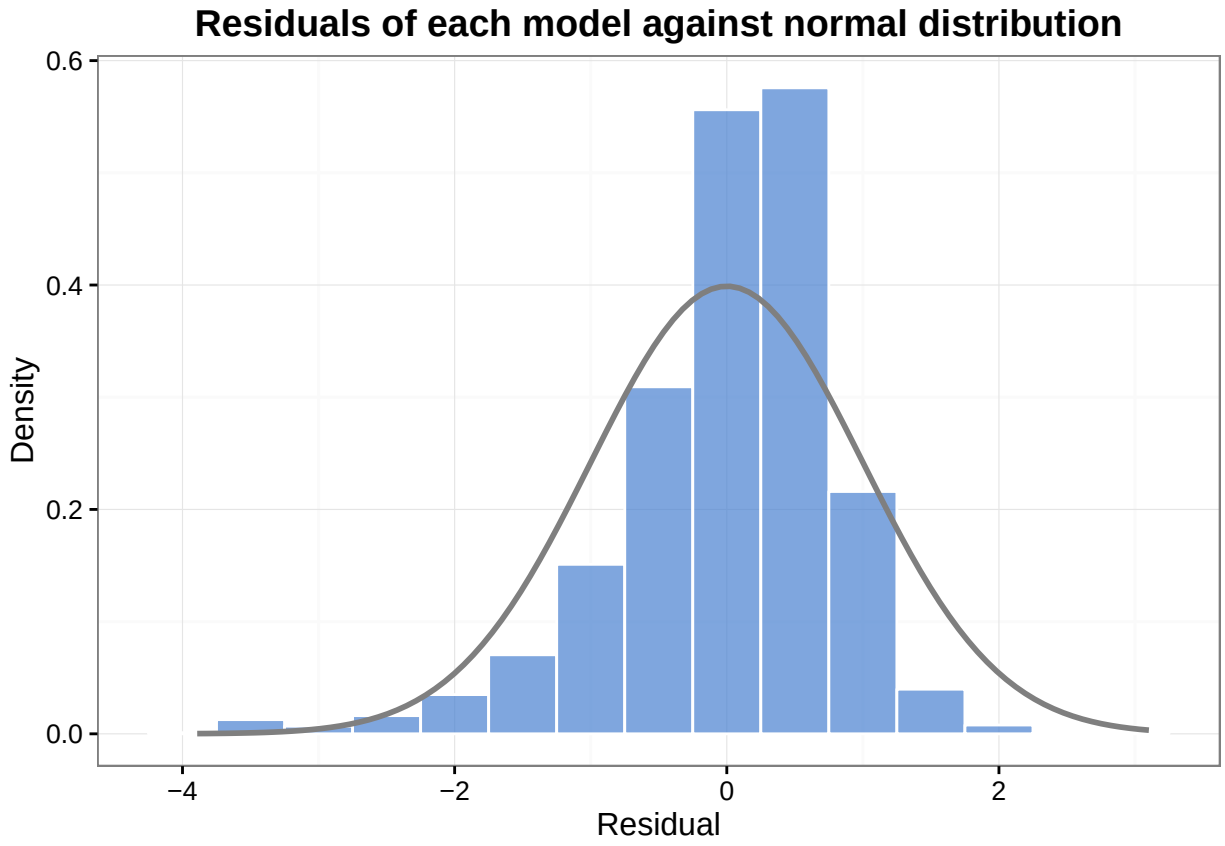

Figure 10: blank

The residuals look more normally distributed for model 3 and model 4, but models 1 and 2 look a little more spread out. We don't have very good resolution of the residuals, but this is mainly because we don't have that much data, only 16 observations, so this is OK.

From here we can see that our plot of the residuals looks pretty normal, and that is a good thing as it is one of the model assumptions! Huzzah!

Homogeneity of errors

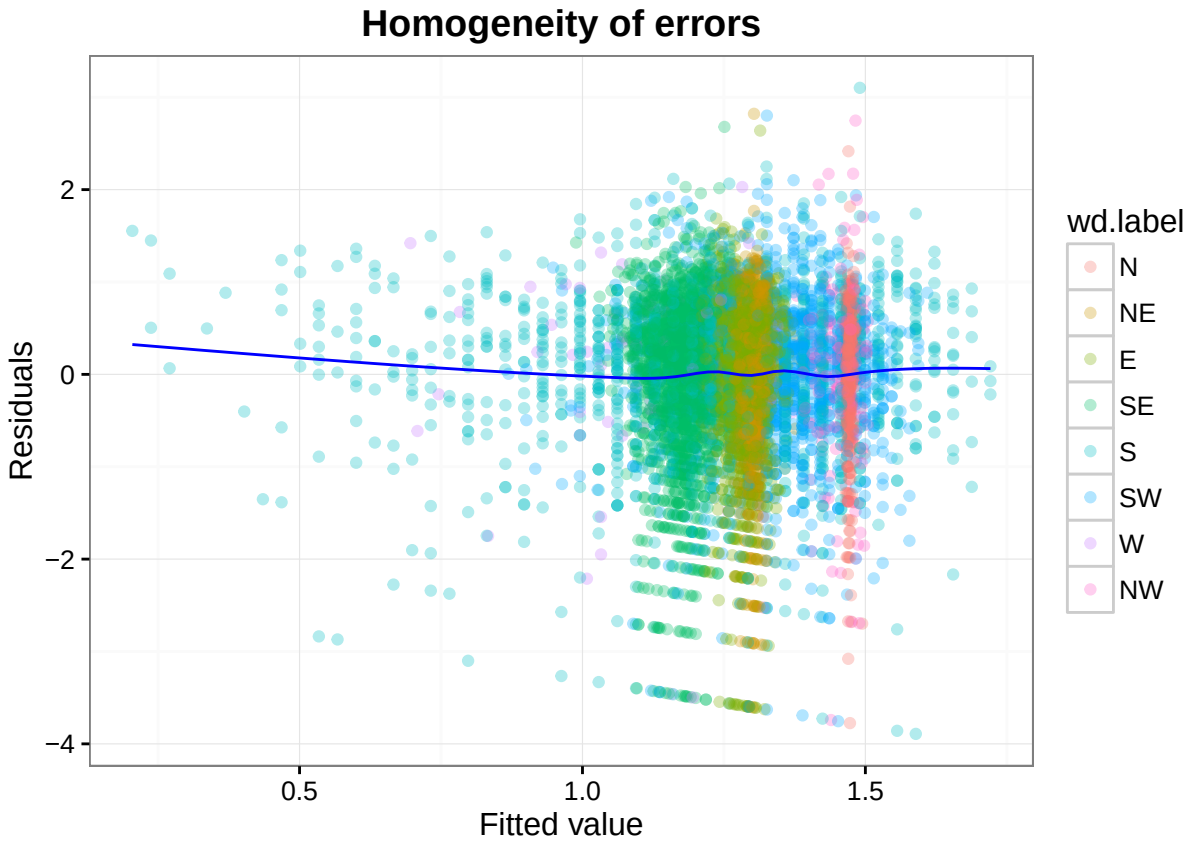

Figure 11: blank

Do the residuals look like they have a mean of zero and constant variance as we move from left to right along the fitted values axis? Do these plots indicate that the residuals are normally distributed and homogenous in their variance?

We can see here that for both models the smaller fitted values (some of which are negative!) have residuals which are all greater than zero (indicating underprediction), that the fitted values in the 25-50 range are being overpredicted (as the residuals are negative) and that for the larger fitted values there is substantially greater variance in the residuals.

QQ plot

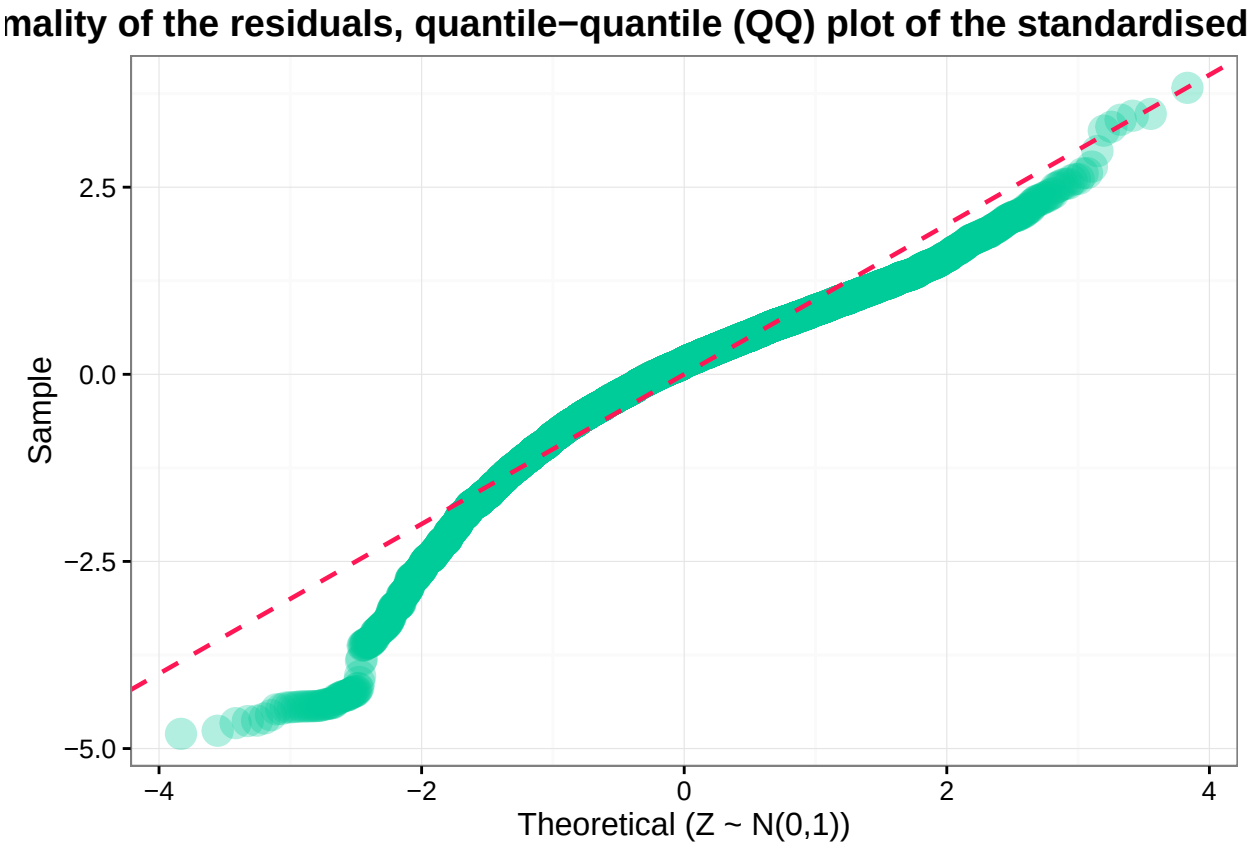

Figure 12: blank

Do these residuals look approximately Normally distributed?

It looks like these residuals are not Normally distributed due to the skewness from the histograms and the departure from the geom_abline() line sample = theoretical. Perhaps there is some structure to the unexplained variation in the residuals.

For model 1., it looks like the model is over-predicting the lower (peak around -15), and under-predicting upper values (points around 100). This suggests that there is some unexplained variation!

Goodness of fit plot

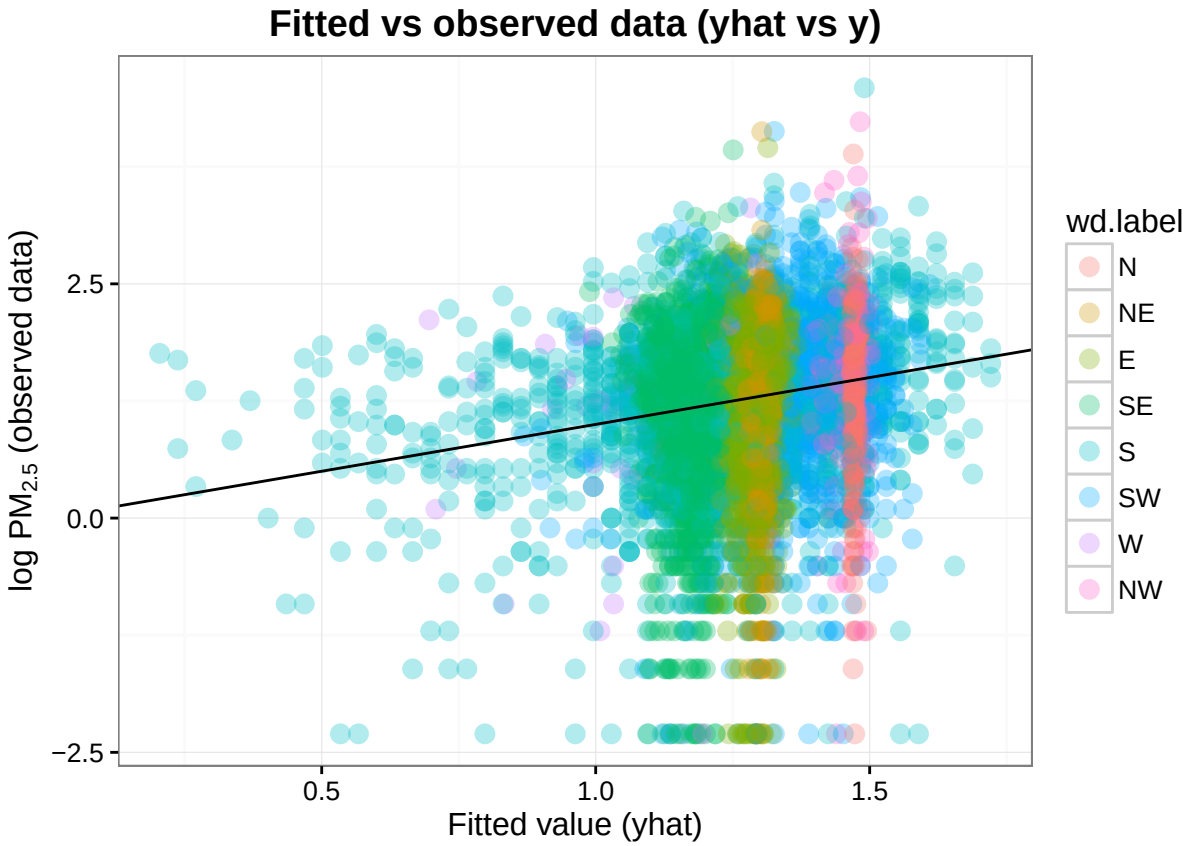

Figure 13: blank

How much do our modelled values, yhat, look like our observed values, y_i Expect them to fall very close to a straight line. Are there any records where the observed $PM_{2.5}$ are either all above or all below where we would expect them to be from the model? (i.e. where the y_i are all bigger or all smaller than the $yhat_i$)

---

5. Interpret

Model interpretation

Link back

Compare

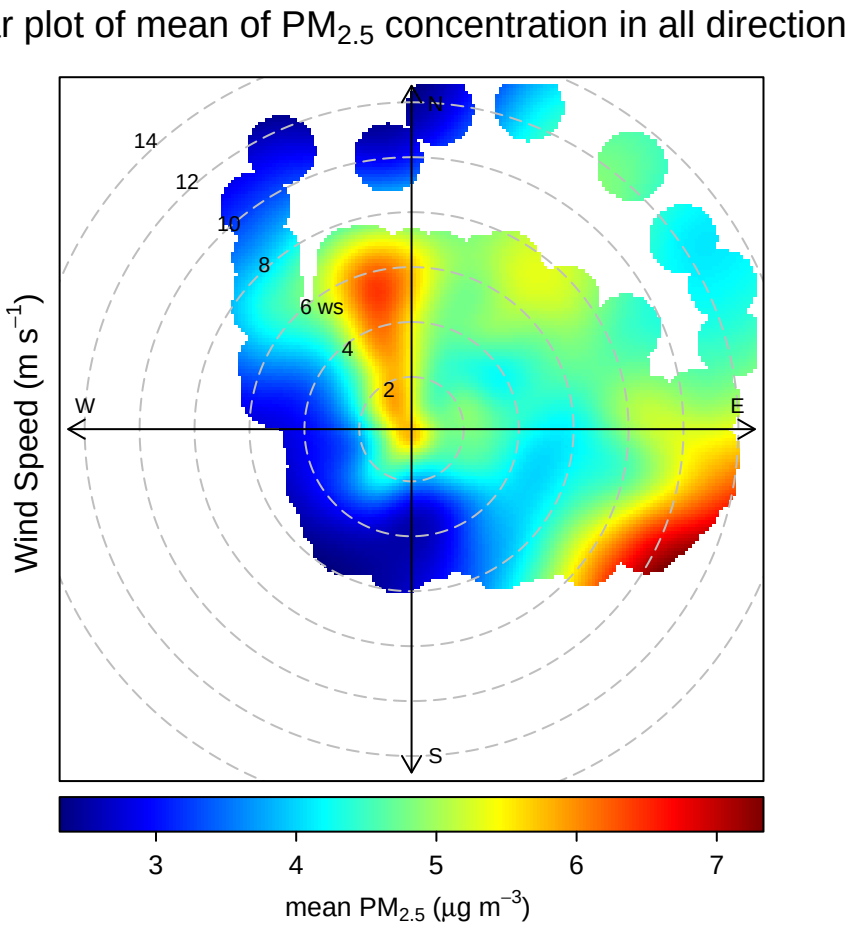

Figure 14: Global average across Oceean Health Index goals.

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

Carslaw D and Ropkins K (2016). (Carslaw & Ropkins, 2012): Open-source tools for the analysis of air pollution data.

References

Carslaw, D. C., & Ropkins, K. (2012). Openair — an r package for air quality data analysis. *Environmental Modelling & Software*, *27–28*(0), 52–61. http://doi.org/10.1016/j.envsoft.2011.09.008

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/