Multivariate Statistics Analysis on Air Quality dataset

Flore Uzan

June 11, 2021

1 Introduction

1.1 Motivation

In recent times, as climate change has become the center of attention for a lot of statistical research, multivariate data sets on air quality have become quite popular for measuring and finding relationships between air pollutants and changes in temperature and humidity. To this end, we decided to work on an air quality data set provided through the University of California, Irvine Machine Learning Repository. To explore the relationship between the predictors and the outcome variables of this data set, we employ techniques such as:

1. Comparison Multivariate Regression and PC Regression
2. Stimulation study on the distribution of the residuals
3. Factor Analysis on the predictors (comparison PC method and ML method)
4. Time Dependent Analysis (exploring how the relationship between outcomes and predictors changes with change in Month and Hour)

1.2 Data

The data set is a multivariate, time-series containing 9358 instances of hourly averaged responses from an array of 5 metal oxide chemical sensors embedded in an Air Quality Chemical Multi-sensor device. Data were recorded from March 2004 to February 2005 (one year) representing the longest freely available recordings of on field deployed air quality chemical sensor devices responses. Ground Truth hourly averaged concentrations for CO, Non Metanic Hydrocarbons, Benzene, Total Nitrogen Oxides (NOx) and Nitrogen Dioxide (NO2) and were provided by a co-located reference certified analyzer. There are 13 attributes in the data set which potentially can be used as predictors with two outcome variables - Relative Humidity (RH) and Absolute Humidity (AH). The attributes are:

1. Date (DD/MM/YYYY)
2. Time (HH.MM.SS)
3. CO(GT): True hourly averaged concentration CO (carbon monoxide) in mg/m^3 (reference analyzer)
4. PT08.S1(CO): PT08.S1 (tin oxide) hourly averaged sensor response (nominally CO targeted)
5. NMHC(GT): True hourly averaged overall NMHC (Non Metanic HydroCarbons) concentration in microg/m^3 (reference analyzer)
6. C6H6(GT): True hourly averaged Benzene concentration in microg/m^3 (reference analyzer)
7. PT08.S2(NMHC): PT08.S2 (titania) hourly averaged sensor response (nominally NMHC targeted)
8. NOx(GT): True hourly averaged NOx (Total Nitrogen Oxides) concentration in ppb (reference analyzer)
9. PT08.S3(NOx): PT08.S3 (tungsten oxide) hourly averaged sensor response (nominally NOx targeted)
10. NO2(GT): True hourly averaged NO2 (Nitrogen Dioxide) concentration in microg/m^3 (reference analyzer)
11. PT08.S4(NO2): PT08.S4 (tungsten oxide) hourly averaged sensor response (nominally NO2 targeted)
12. PT08.S5(O3): PT08.S5 (indium oxide) hourly averaged sensor response (nominally O3 targeted)
13. T: Temperature in Â°C
14. RH: Relative Humidity (%)
15. AH: Absolute Humidity

2 Exploratory Data Analysis

2.1 Missing values

Missing values in this data set were tagged with a -200 value. After identifying them, we decided to delete the variable 'PT08.S3(NOx)' that recorded 8443 missing values. Furthermore, 366 values were missing for multiple rows in the data set and hence we decided to remove them. We finally had a total of 8991 observation with 'Date' and 'Time' for each in addition to the 12 responses from the array of 5 metal oxide chemical sensors.

The remaining missing values in a few of the columns were imputed with the mean of the column, using the 'na.aggregate()' function, so as to minimize the variance caused by them.

2.2 Correlation

We can observe high correlation above 0.80 among some of the pollutants: CO(GT) - Concentrations of carbon monoxide, PTO8.S1(CO) - Tin oxide, C6H6(GT), Benzene, PTO8.S2(NMHC), Titanium and PTO8.S5(O3). In the same way, PTO8.S5(O3) - Indium oxide is negatively and highly correlated with the rest of the pollutants. The Nitrogen molecules - NOx(GT) and NO2(GT) also have a strong correlation equal to 0.76. Looking at the two outcome variables, they are surprisingly not highly correlated. However, RH is negatively correlated with Temperature while AH is positively correlated with it. For the month and hour of the day, we do not observe a strong correlation.
We explore the time dependent relationship among these variables separately in section 6 of this report.

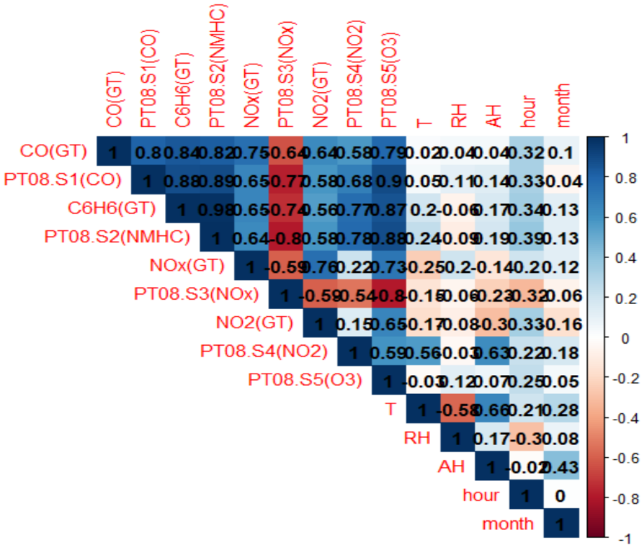


Figure 1: Correlation Plot

2.3 Distribution of Each Variable

Figure 2 presents the density plots of each of our variables. Looking at the density of each variables, we do not observe any particular patterns in our data however there are a few important points that we can take note of:

1. The variable 'NO2.GT' and 'NOx.GT' have density plots which look quite sharp and have high variance. Similarly, the variable 'CO.GT' also has

quite a sharp density plot with high variance. This seems to suggest that these three variables could perhaps be factored together.

2. Both our outcome variables have quite an even spread and look fairly symmetrical. The variable AH though has two small peaks.

3. Variables like 'PT08.S1.CO', 'PT08.S2.NMHC', 'PT08.S3.NOx' and 'PT08.S5.O3' also have very similar density plots suggesting that they perhaps could be factored together.

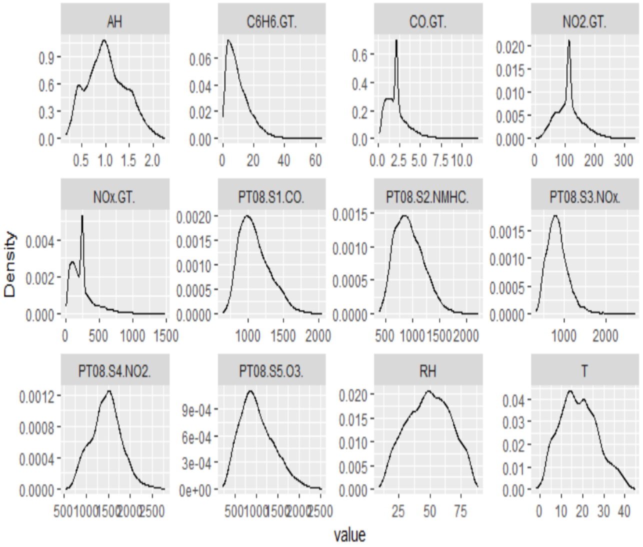


Figure 2: Density plots

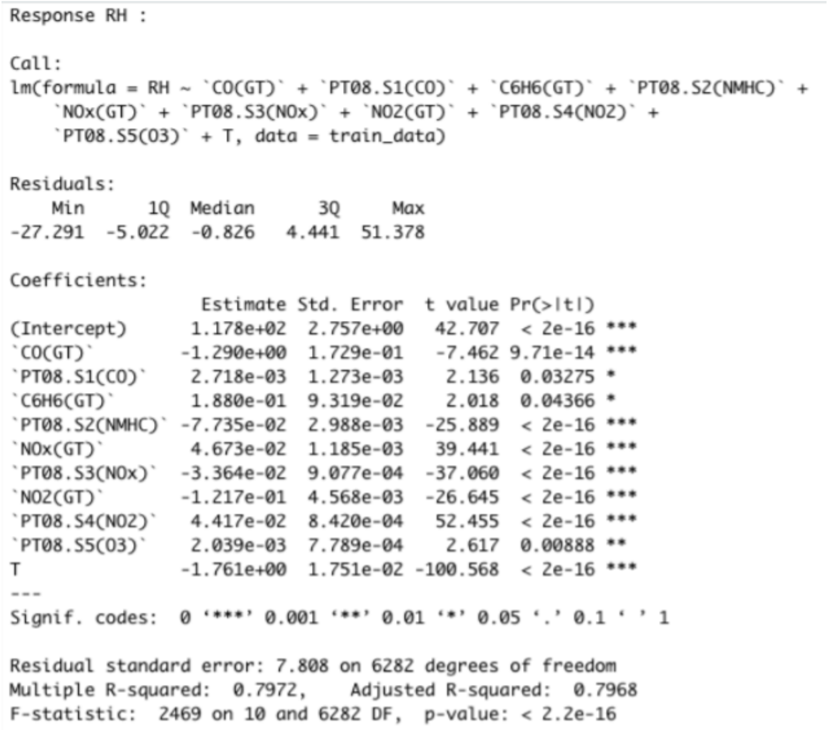
3 Multivariate vs PCA Regression

3.1 Multivariate Regression

As described in the introduction section, our data set provided us with 13 potential attributes which could be used as predictors. Upon further investigation, we dropped the 5th attribute - NMHC(GT) on account of a large number of missing values. So we were finally left with 12 potential predictors and 2 outcomes - Relative Humidity (RH) and Absolute Humidity (AH). For this analysis, we decide drop the date and the time and keep the 10 other predictors.

As a first step in the model fitting process, we split our data into training and testing. We decided to train our model on 70% of the total data and then test our model on the remaining 30% of the data. This implied that our training set had 6293 observations on the two outcome variables and our test set had 2698 observations on the two outcome variables.

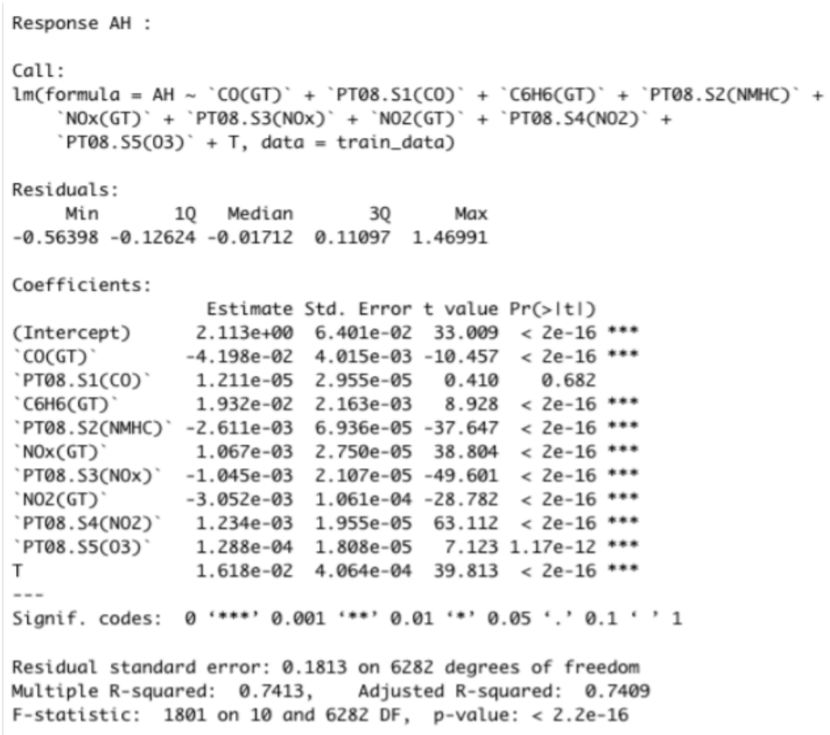
We then ran a multivariate regression model on our two outcome variables jointly to get an R output for each of our two outcome variables.
In Figure 3, we see the R output for our outcome variable - Relative Humidity (RH). As is visible in the figure, all of our predictors seem significant for this outcome variable. With an R^2 value of close to 80% and the F-statistic p-value almost close to 0, this model seems a very decent fit.
In Figure 4, we see the R output for our outcome variable - Absolute Humidity (AH). In this case, as is visible in the figure, all of our predictors except 'PT08.S1' which is a measure of 'Tin Oxide' seem significant for this outcome variable. However, even in such a case, with an R^2 value of close to 75% and the F-statistic p-value almost close to 0, this model too seems a very decent fit.



```
Response RH :

Call:
lm(formula = RH ~ `CO(GT)` + `PT08.S1(CO)` + `C6H6(GT)` + `PT08.S2(NMHC)` +
    `NOx(GT)` + `PT08.S3(NOx)` + `NO2(GT)` + `PT08.S4(NO2)` +
    `PT08.S5(O3)` + T, data = train_data)

Residuals:
    Min      1Q  Median      3Q     Max
-27.291  -5.022  -0.826   4.441  51.378

Coefficients:
                  Estimate Std. Error  t value Pr(>|t|)
(Intercept)      1.178e+02  2.757e+00   42.707  < 2e-16 ***
`CO(GT)`        -1.290e+00  1.729e-01   -7.462 9.71e-14 ***
`PT08.S1(CO)`    2.718e-03  1.273e-03    2.136  0.03275 *
`C6H6(GT)`       1.880e-01  9.319e-02    2.018  0.04366 *
`PT08.S2(NMHC)` -7.735e-02  2.988e-03  -25.889  < 2e-16 ***
`NOx(GT)`        4.673e-02  1.185e-03   39.441  < 2e-16 ***
`PT08.S3(NOx)`  -3.364e-02  9.077e-04  -37.060  < 2e-16 ***
`NO2(GT)`       -1.217e-01  4.568e-03  -26.645  < 2e-16 ***
`PT08.S4(NO2)`   4.417e-02  8.420e-04   52.455  < 2e-16 ***
`PT08.S5(O3)`    2.039e-03  7.789e-04    2.617  0.00888 **
T               -1.761e+00  1.751e-02 -100.568  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.808 on 6282 degrees of freedom
Multiple R-squared:  0.7972,    Adjusted R-squared:  0.7968
F-statistic:  2469 on 10 and 6282 DF,  p-value: < 2.2e-16
```

Figure 3: R Output for response RH

The multivariate regression model fit is further emboldened by the Root Mean Squared Error (RMSE) values for both outcome variables on our remaining 30% test data. For RH, the RMSE is a value of 11.352 and for AH, the RMSE is a value 0.170. Accounting for the different scales of these variables,



```
Response AH :

Call:
lm(formula = AH ~ `CO(GT)` + `PT08.S1(CO)` + `C6H6(GT)` + `PT08.S2(NMHC)` +
    `NOx(GT)` + `PT08.S3(NOx)` + `NO2(GT)` + `PT08.S4(NO2)` +
    `PT08.S5(O3)` + T, data = train_data)

Residuals:
     Min       1Q   Median       3Q      Max
-0.56398 -0.12624 -0.01712  0.11097  1.46991

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)      2.113e+00  6.401e-02  33.009  < 2e-16 ***
`CO(GT)`        -4.198e-02  4.015e-03 -10.457  < 2e-16 ***
`PT08.S1(CO)`    1.211e-05  2.955e-05   0.410    0.682
`C6H6(GT)`       1.932e-02  2.163e-03   8.928  < 2e-16 ***
`PT08.S2(NMHC)` -2.611e-03  6.936e-05 -37.647  < 2e-16 ***
`NOx(GT)`        1.067e-03  2.750e-05  38.804  < 2e-16 ***
`PT08.S3(NOx)`  -1.045e-03  2.107e-05 -49.601  < 2e-16 ***
`NO2(GT)`       -3.052e-03  1.061e-04 -28.782  < 2e-16 ***
`PT08.S4(NO2)`   1.234e-03  1.955e-05  63.112  < 2e-16 ***
`PT08.S5(O3)`    1.288e-04  1.808e-05   7.123 1.17e-12 ***
T                1.618e-02  4.064e-04  39.813  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1813 on 6282 degrees of freedom
Multiple R-squared:  0.7413,    Adjusted R-squared:  0.7409
F-statistic:  1801 on 10 and 6282 DF,  p-value: < 2.2e-16
```

Figure 4: R Output for response AH

both RMSE values seem to be fairly low indicating our multivariate regression model generalizes well on test data as well.

To round up or discussion on multivariate regression for this data set, we looked at the histogram of residuals for both outcome variables. From Figure **??**, we observed that the residuals for the outcome variable RH seem slightly skewed to the right. And similarly, from Figure **??**, we observed that the residuals for the outcome variable AH also seem slightly skewed to the right. This suggested that one of the fundamental assumptions of a multivariate regression model - normality of residuals - might not be met in our case. This implied that our model parameter estimates might be slightly biased and hence our model may not be as useful as it seemed looking at the R output earlier. In Section 4 of this report, we will describe a small simulation study that we conducted in order to investigate skew-normal residuals and their effects on parameter estimates of a simple linear regression model.

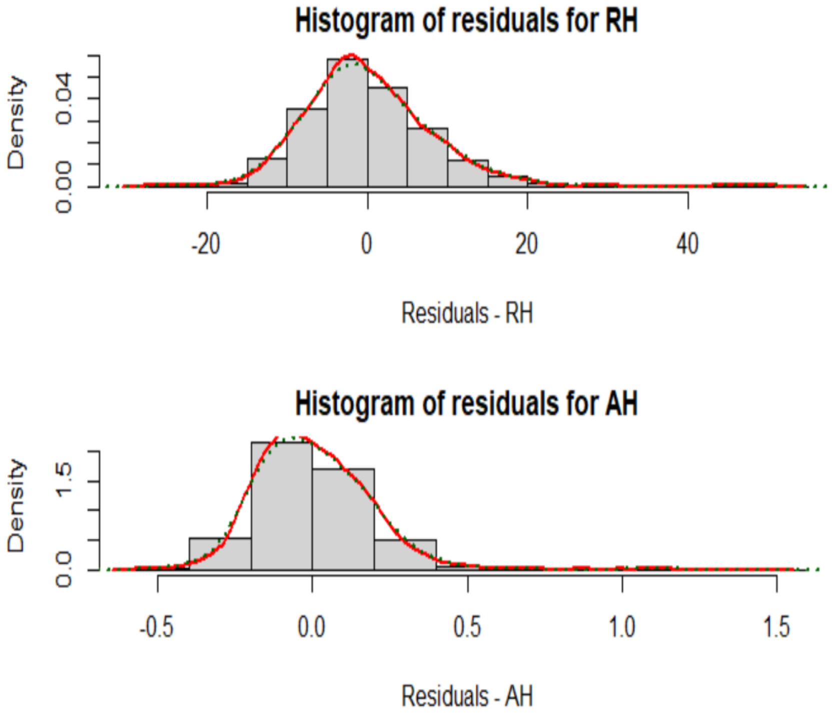


Figure 5: Histogram of residuals for outcome RH and AH

3.2 PCA Regression

We then moved onto conducting a principal component analysis (PCA) of our training data to figure out if we can summarize its variability in a small number of components. Once we were able to achieve such a summary, we used these principal components as predictors in a linear regression model to conduct principal component regression analysis (PCR). We also compared the RMSE obtained on our test set from the multivariate regression model to the RMSE on the test set that we obtained via PCR.

The first step is to determine the number of PCs that we will use in our analysis. We look at the scree plot giving the percentage of variance explained for each added PC. To determine the appropriate number of components, we look for an elbow (bend) in the scree plot. The number of components is taken to be the point at which the variance explained by the remaining PCs are relatively small and all about the same size. From Figure 6, we observed a fairly sharp elbow in the scree plot when we have 3 principal components and a fairly constant linear line after that. This suggested using 3 components for our PCR analysis might suffice. Further, we also looked at the PCA output on our training data from R. From Figure 7, we observed that the first three principal components

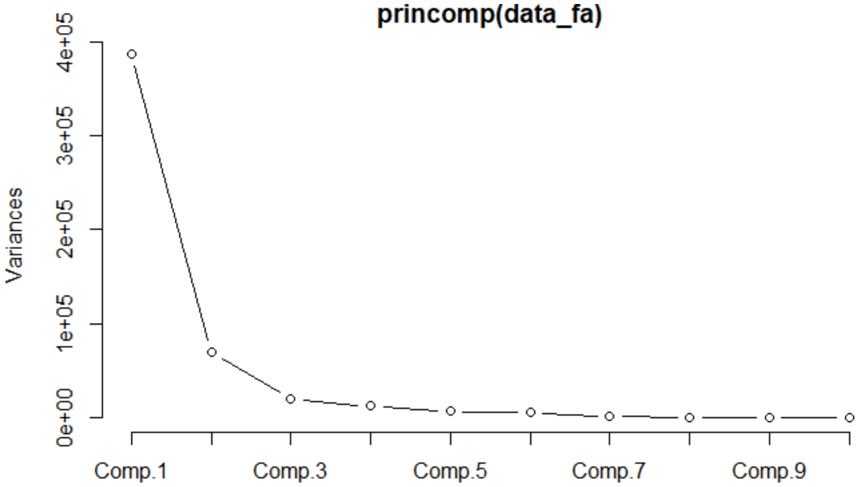
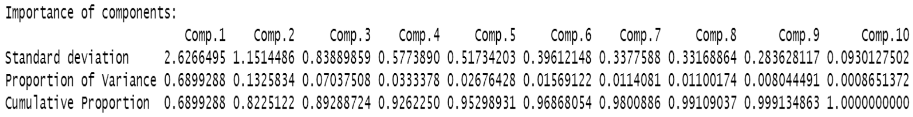


Figure 6: Screeplot of the principal components

account for about 90% of the total variance in our training data which also seemed to suggest that using 3 principal components was the right choice.



```
Importance of components:
                          Comp.1    Comp.2     Comp.3    Comp.4     Comp.5     Comp.6    Comp.7     Comp.8      Comp.9      Comp.10
Standard deviation     2.6266495 1.1514486 0.83889859 0.5773890 0.51734203 0.39612148 0.3377588 0.33168864 0.283628117 0.0930127502
Proportion of Variance 0.6899288 0.1325834 0.07037508 0.0333378 0.02676428 0.01569122 0.0114081 0.01100174 0.008044491 0.0008651372
Cumulative Proportion  0.6899288 0.8225122 0.89288724 0.9262250 0.95298931 0.96868054 0.9800886 0.99109037 0.999134863 1.0000000000
```

Figure 7: PCA output from R

We used the 'pcr()' function in R to fit a principal component regression to our model and then used this PCR model to make predictions on our test set. We recorded the RMSE for our first outcome variable - RH to be 16.684 and for our second outcome variable - AH to be 0.243. These are not very far away from what we observed in the multivariate regression case but here we have only used 3 principal components which provided us with computational efficiency and summarized our data more efficiently.

In Table 1, we summarize our RMSE results from the multivariate regression model and the PCR model with number of components being 3.

Outcome	Multi. Regression	PCR with ncomp = 3
RH	11.352	16.684
AH	0.170	0.243

Table 1: Table for RMSE from Multivariate Regression and PCA

4 Simulation Study

In section 3, we had concluded from looking at the histogram of residuals from both the outcome variables that our multivariate regression model estimates maybe biased. This was because the residuals were slightly-skewed to the right which violated the assumption of normality of residuals. In this section, we present results from a small simulation study that we conducted to investigate how the skewness of the residuals affects the parameter estimates of a linear regression model.

4.1 Setup of Simulation Study

We consider a case of simple linear regression and hence we have our true model of the form:

$Y_{RH} = \beta_{0,RH} + \beta_{1,RH}X + \beta_{2,RH}X + \beta_{3,RH}X + \beta_{4,RH}X + \beta_{5,RH}X + \beta_{6,RH}X + \beta_{7,RH}X + \beta_{8,RH}X + \beta_{9,RH}X + \beta_{10,RH}X + \epsilon$

$Y_{AH} = \beta_{0,AH} + \beta_{1,AH}X + \beta_{2,AH}X + \beta_{3,AH}X + \beta_{4,AH}X + \beta_{5,AH}X + \beta_{6,AH}X + \beta_{7,AH}X + \beta_{8,AH}X + \beta_{9,AH}X + \beta_{10,AH}X + \epsilon$

where the betas represent the 10 estimators from the multivariate regression, that we consider to be the true parameter value.

For the data, we use a the matrix called X composed of 8991 rows for each observations and 11 columns including the 10 predictors and the first column of 1 for the intercept.

We generate two separate form of errors to compute our outcome variable, Y with the same mean matrix and covariance matrix using the function *cp2dp* from the library *sn* :

- $\epsilon \sim$ Multivariate Normal(μ, Σ)
- $\epsilon \sim$ Skew Normal(μ, Σ, γ)

where $\mu = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$, $\Sigma = \begin{pmatrix} cov(RH) & cor(RH, AH) \\ cor(RH, AH) & cov(AH) \end{pmatrix}$ and $\gamma = \begin{pmatrix} skew(RH)/2 \\ skew(AH)/2 \end{pmatrix}$.

Finally, we fit a linear model $Y \sim X$ and compute our parameter estimates. $\hat{\beta}_{j,RH}$ and $\hat{\beta}_{j,AH}$ for j = 0, ..., 10. We set the number of simulation iterations, $nsim = 1000$.

4.2 Results of Simulation Study

With the simulation setup as described above, we obtained the following sampling distribution of each $\hat{\beta}$s for RH generated from Normal errors in green and Skew Normal errors in blue in Figures 8 and 9. As is evident from the plot, it is centered very close to the true value of the parameters represented by the vertical line and has a symmetrical shape very close to a normal distribution. We do not observe any difference between of variance or strong bias between the two distributions.

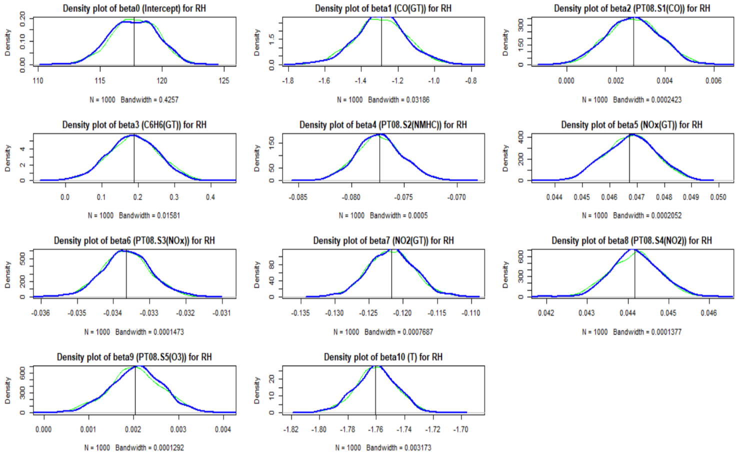


Figure 8: Sampling distribution of the βs for RH: Normal Errors (green) and Skew Normal Errors (blue)

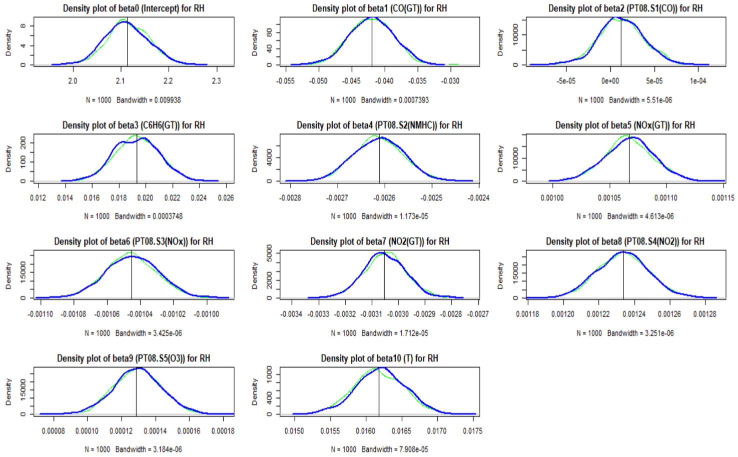
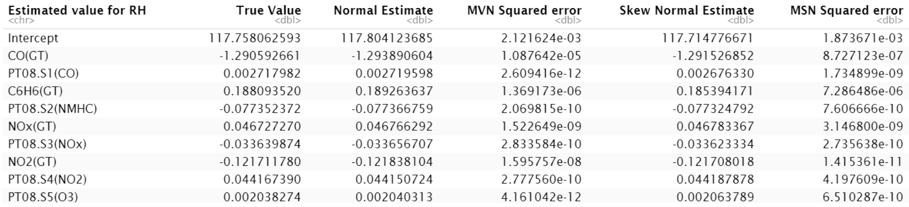


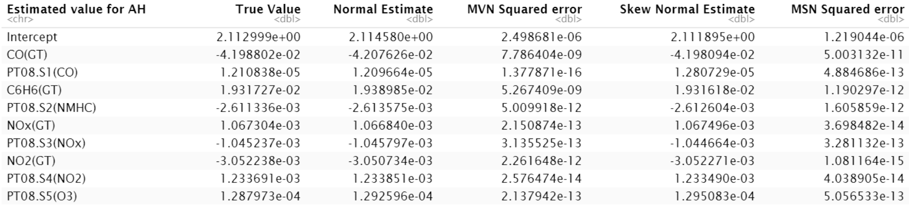
Figure 9: Sampling distribution of the βs for AH: Normal Errors (green) and Skew Normal Errors (blue)

In the tables in Figures 10 and 11, the normal and skew-normal errors provided very close estimates for the predictrs. MVN Squared error is the squared difference between the true value and the Normal estimate and the MSN Squared error is the squared difference between the true value and the Skewed Normal estimate.



Estimated value for RH	True Value	Normal Estimate	MVN Squared error	Skew Normal Estimate	MSN Squared error
<chr>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
Intercept	117.758062593	117.804123685	2.121624e-03	117.714776671	1.873671e-03
CO(GT)	-1.290592661	-1.293890604	1.087642e-05	-1.291526852	8.727123e-07
PT08.S1(CO)	0.002717982	0.002719598	2.609416e-12	0.002676330	1.734899e-09
C6H6(GT)	0.188093520	0.189263637	1.369173e-06	0.185394171	7.286486e-06
PT08.S2(NMHC)	-0.077352372	-0.077366759	2.069815e-10	-0.077324792	7.606666e-10
NOx(GT)	0.046727270	0.046766292	1.522649e-09	0.046783367	3.146800e-09
PT08.S3(NOx)	-0.033639874	-0.033656707	2.833584e-10	-0.033623334	2.735638e-10
NO2(GT)	-0.121711780	-0.121838104	1.595757e-08	-0.121708018	1.415361e-11
PT08.S4(NO2)	0.044167390	0.044150724	2.777560e-10	0.044187878	4.197609e-10
PT08.S5(O3)	0.002038274	0.002040313	4.161042e-12	0.002063789	6.510287e-10

Figure 10: Table for parameter estimates from Normal and Skew-Normal Errors for RH



Estimated value for AH	True Value	Normal Estimate	MVN Squared error	Skew Normal Estimate	MSN Squared error
<chr>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
Intercept	2.112999e+00	2.114580e+00	2.498681e-06	2.111895e+00	1.219044e-06
CO(GT)	-4.198802e-02	-4.207626e-02	7.786404e-09	-4.198094e-02	5.003132e-11
PT08.S1(CO)	1.210838e-05	1.209664e-05	1.377871e-16	1.280729e-05	4.884686e-13
C6H6(GT)	1.931727e-02	1.938985e-02	5.267409e-09	1.931618e-02	1.190297e-12
PT08.S2(NMHC)	-2.611336e-03	-2.613575e-03	5.009918e-12	-2.612604e-03	1.605859e-12
NOx(GT)	1.067304e-03	1.066840e-03	2.150874e-13	1.067496e-03	3.698482e-14
PT08.S3(NOx)	-1.045237e-03	-1.045797e-03	3.135525e-13	-1.044664e-03	3.281132e-13
NO2(GT)	-3.052238e-03	-3.050734e-03	2.261648e-12	-3.052271e-03	1.081164e-15
PT08.S4(NO2)	1.233691e-03	1.233851e-03	2.576474e-14	1.233490e-03	4.038905e-14
PT08.S5(O3)	1.287973e-04	1.292596e-04	2.137942e-13	1.295083e-04	5.056533e-13

Figure 11: Table for parameter estimates from Normal and Skew-Normal Errors for AH

Overall, the results from the simulation study have helped us conclude that the skew-normal distributed errors are as good (or even better) as the normal distributed errors. Therefore, treating the residuals as Skew-Normal distributed would be a good choice for our study. We can also conclude that, even though we had skewed residuals, the estimates with a normal distributed errors from our multivariate regression model were robust. Therefore,

5 Factor Analysis

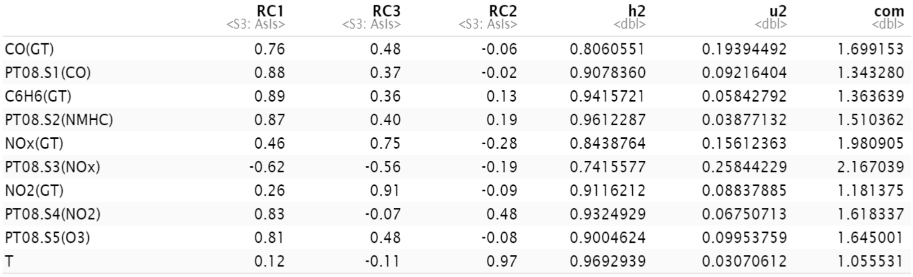
We perform a factor analysis using the 10 predictors of the Relative Humidity and the Absolute Humidity using the principal component analysis and the maximum likelihood method.

The correlation matrix presented in the Exploratory data Analysis section presents strong correlations between the explanatory variables. It implies that a small number of factor might be enough for a factor analysis.

5.1 Principal Component Method

In section 3.2, we determined the number m of PCs that we should use in the PC regression analysis. We could see that about 90% of the cumulative variance was explained by the three first PCs. We look more in depth at the three PCs in order to propose an interpretation of the three principal components.

From the Figure 12, it is clear that variables 1, 2, 3, 4, 6, 8 and 9 define factor



	RC1	RC3	RC2	h2	u2	com
	<S3: AsIs>	<S3: AsIs>	<S3: AsIs>	<dbl>	<dbl>	<dbl>
CO(GT)	0.76	0.48	-0.06	0.8060551	0.19394492	1.699153
PT08.S1(CO)	0.88	0.37	-0.02	0.9078360	0.09216404	1.343280
C6H6(GT)	0.89	0.36	0.13	0.9415721	0.05842792	1.363639
PT08.S2(NMHC)	0.87	0.40	0.19	0.9612287	0.03877132	1.510362
NOx(GT)	0.46	0.75	-0.28	0.8438764	0.15612363	1.980905
PT08.S3(NOx)	-0.62	-0.56	-0.19	0.7415577	0.25844229	2.167039
NO2(GT)	0.26	0.91	-0.09	0.9116212	0.08837885	1.181375
PT08.S4(NO2)	0.83	-0.07	0.48	0.9324929	0.06750713	1.618337
PT08.S5(O3)	0.81	0.48	-0.08	0.9004624	0.09953759	1.645001
T	0.12	-0.11	0.97	0.9692939	0.03070612	1.055531

Figure 12: R Output for PC method with rotation Varimax

1 with high loadings on factor 1 and small or negligible loadings on factor 2 and 3. The factor 1 is a explained by CO(GT), concentrations of carbon monoxide, C6H6(GT), the Benzene, PTO8.S1(CO), tin oxide, PTO8.S2(NMHC), titanium, PTO8.S3(NOx) and PTO8.S4(NO2), tungsten oxide and PTO8.S5(O3), indium oxide. Perhaps we can call this the "pollutants" factor.

Variable 10 define factor 2 with high loadings on factor 2 and small or negligible loadings on factor 1 and 3.

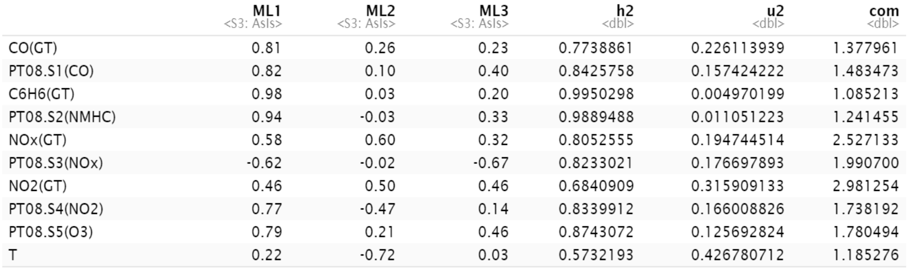
The factor 2 is only defined by T, the temperature. Perhaps, we can call it the "temperature" factor.

Variables 5 and 7 define factor 3 with high loadings on factor 3 and small or negligible loadings on factor 1 and 2.

The factor 3 is explained by: NOx(GT), the Total Nitrogen Oxides and NO2(GT), the Nitrogen Dioxide. Perhaps we can call this the "Nitrogen" factor.

5.2 Maximum Likelihood Method

In the same way, we look at the cumulative variance explained, equal to 82% for $m = 3$ and the cumulative proportion explained by the maximum likelihood method which is equal to 100% after $m = 3$.



	ML1	ML2	ML3	h2	u2	com
	<S3: AsIs>	<S3: AsIs>	<S3: AsIs>	<dbl>	<dbl>	<dbl>
CO(GT)	0.81	0.26	0.23	0.7738861	0.226113939	1.377961
PT08.S1(CO)	0.82	0.10	0.40	0.8425758	0.157424222	1.483473
C6H6(GT)	0.98	0.03	0.20	0.9950298	0.004970199	1.085213
PT08.S2(NMHC)	0.94	-0.03	0.33	0.9889488	0.011051223	1.241455
NOx(GT)	0.58	0.60	0.32	0.8052555	0.194744514	2.527133
PT08.S3(NOx)	-0.62	-0.02	-0.67	0.8233021	0.176697893	1.990700
NO2(GT)	0.46	0.50	0.46	0.6840909	0.315909133	2.981254
PT08.S4(NO2)	0.77	-0.47	0.14	0.8339912	0.166008826	1.738192
PT08.S5(O3)	0.79	0.21	0.46	0.8743072	0.125692824	1.780494
T	0.22	-0.72	0.03	0.5732193	0.426780712	1.185276

Figure 13: R Output for ML method with rotation Varimax

It is clear from the Figure 13 that the 3 factors are explained by the same variables as before.

The factor 1 is a explained by CO(GT), concentrations of carbon monoxide, C6H6(GT), the Benzene, PTO8.S1(CO), tin oxide, PTO8.S2(NMHC), titanium, PTO8.S3(NOx) and PTO8.S4(NO2), tungsten oxide and PTO8.S5(O3), indium oxide potentially called the “pollutants” factor.

The factor 2 is only defined by T, the temperature that could be called the “temperature” factor.

The factor 3 is explained by: NOx(GT), the Total Nitrogen Oxides and NO2(GT), the Nitrogen Dioxide, could be called the “Nitrogen” factor.

6 Time Dependent Analysis

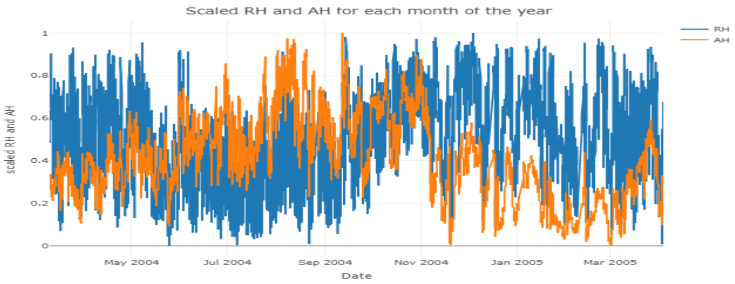


Figure 14: Scaled value of RH and AH over the year

As a preliminary overview, we plot each variable over the year. In order to compare RH and AH together, we standardize the two variables by subtracting the mean and diving by the standard deviation. As shown on the Figure 14, RH first decreases and then increases, inversely, AH first increases and then increases.

To identify any potential trend, we divide the study into two parts: a first analysis per month of the year and a second per hour of the day.

6.1 Month

The boxplots in Figures 15 and 16 show a clear trend in our data. RH drops around July while AH peaks around August. It confirms the possibility of a relationships between the month and the relative humidity and the absolute humidity.

We then aggregate the mean RH and AH for each month and study any potential pattern or trend among the months in the Figure 17. We standardize both value to plot the histograms of RH and AH together. Again, we can

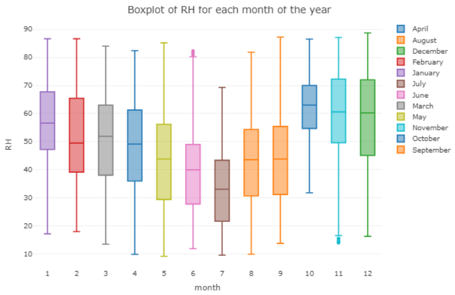


Figure 15: RH by month

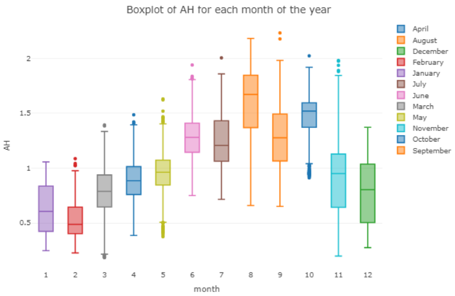


Figure 16: AH by month

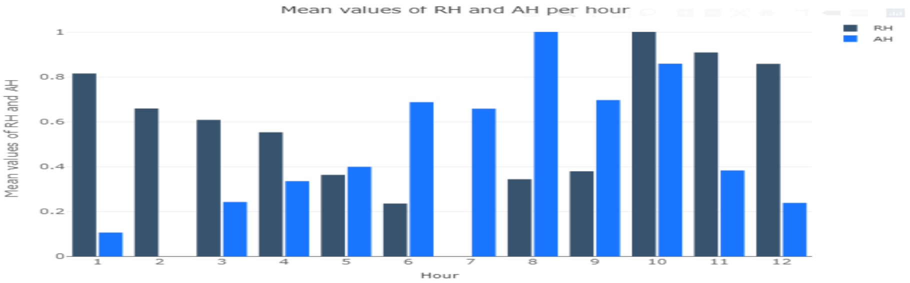


Figure 17: Scaled value of RH and AH over month

observe a trend in the data: RH decreases around August/September and AH increases around August.

We realize a change points or breakpoints analysis to determine abrupt variations in our time series data that could represent transitions between different states. In this case, the change points detection algorithms will help us to detect any relationship between the time of the day and the variation of the RH and AH.

We will compare a few change point detection method available in R given different time series dynamics and research questions.

changepoint

The first technique uses the package *changepoint*. It is focused on intercept-only changes. It can estimate changes in means (cpt.mean), variance (cpt.var), or both (cpt.meanvar). It can recover ML estimates of the intercepts. It does not estimate uncertainty, nor model checking. It only takes a response variable, so the change point is the data index, not the point on an x-axis. In the Figure 18, it detects a mean of 43.06282, and a change of mean after "2004-09-29" equal to 56.08791. In the Figure 19, the change point is at "2004-11-07 with a mean of 1.2160264 and then 0.7038403 for AH. We can see that the trends are reversed.

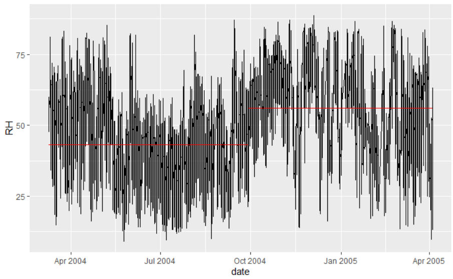


Figure 18: Changepoint in RH

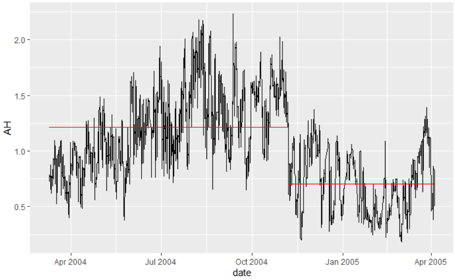


Figure 19: Changepoint in AH

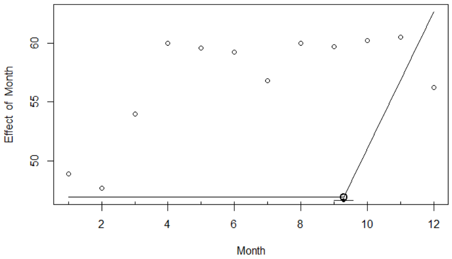


Figure 20: Segments in RH

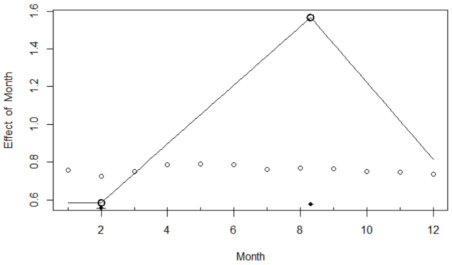


Figure 21: Segments in AH

segmented

The second package is called *segmented.* The segmented package provides functions for segmented or broken-line models, which are regression models where the relationships between the response and one or more explanatory variables are piece wise linear. The number of breakpoints of each segmented relationship must be a-priori specified. In our case, the algorithm detects two segments in RH (Figure 20) with a change in September and three segments with changes in February and August (Figure 21) .

tree

The package *tree*-based methods for regression and classification uses an

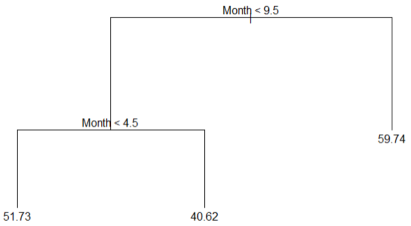


Figure 22: Tree for RH

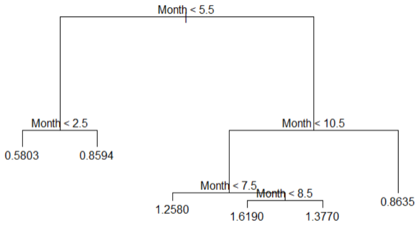


Figure 23: Tree for AH

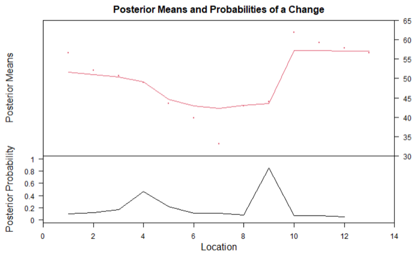


Figure 24: Bcp in RH

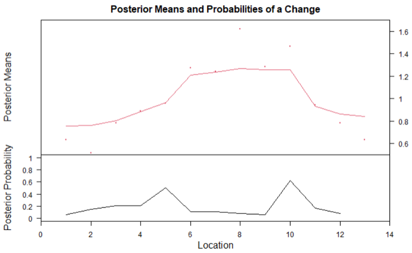


Figure 25: Bcp in AH

alternative approach. It involves stratifying or segmenting the predictor space into a number of simple regions. The value at which the regions are split can also be seen as change points in the predictor. While there are numerous tree methods (e.g. boosting, bagging, random forest) and implementations in R we will here use the simple single decision tree approach that is provided by the tree package. Again, RH identifies a change point in September (Figure 22). However, the main change point for AH is May when using the tree algorithm (Figure 23).

bcp

The last approach uses the *bcp* package designed to perform Bayesian single change point analysis of univariate time series. It returns the posterior probability of a change point occurring at each time index in the series. This package is only designed to detect changes in the mean of independent Gaussian observations with its core function bcp(). For RH, the lower posterior probability plot shows that at the locations 9 and 21, the probability of a change is very high. We can get the exact locations where probabilities are high (e.g. $> 70\%$) and we have a probability of .9 the points 9 that represents September (Figure 24). However for AH, none of the month have a probability of changing point higher than 70% (Figure 25).

Overall, this list of change point detection methods is surely not exclusive but gives us an idea of where RH and AH change depending on the month. For RH, the four methods identify a change point in September, while for AH the responses are diverging. The changepoint method identifies November, the segmented package identifies change point in February and August, the tree identifies a change point in May and the Bayesian approach does not identify any of them. If we look at the boxplot and the histogram of AH, AH has higher values in the Fall which corresponds to the month identified with our detection methods.

6.2 Hour

We proceed to the same analysis for the hour of the day.

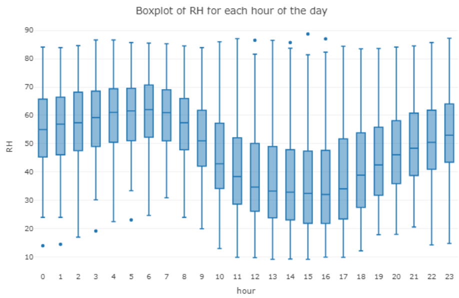


Figure 26: RH by hour

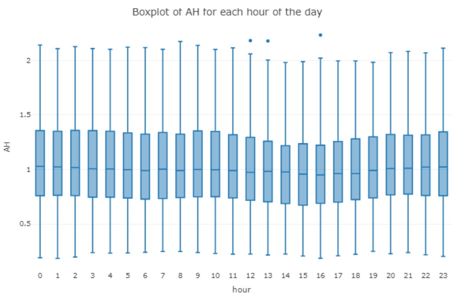


Figure 27: AH by hour

The boxplots in Figures 26 and 27 of RH presents a clear trend in the values of RH, with a drop around 2pm while AH does not present any clear dependence to the time of the day.

We then aggregate the mean of RH and AH for each hour of the experiment and we plot the standardized RH and AH in a histogram in Figure 28. This time, we observe a drop in the values of RH and AH at 2pm.

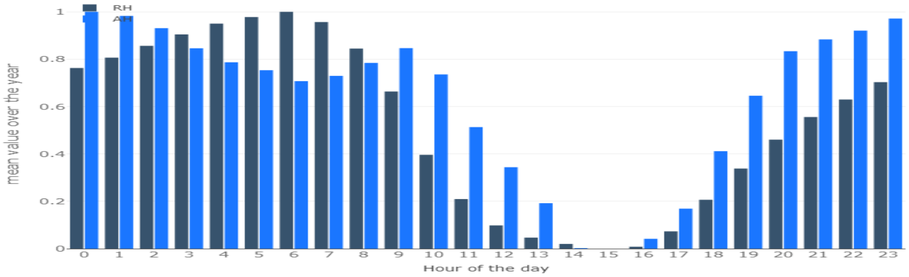


Figure 28: Scaled value of RH and AH by hour

In the same way, we proceed to our change point analysis.

changepoint

First, the *changepoint* package suggests that we only have one change point with RH at 10am (Figure 29) and no one with AH (Figure 30).

segmented

The *segmented* package gives us 2 similar change points for each variables. The change points for RH are 7am and 2pm (Figure 31) and are 9am and 2pm for AH (Figure 32). These values correspond to the trend observe on the histogram.

tree

When using the *tree* package, we have 2 different change points around 9am

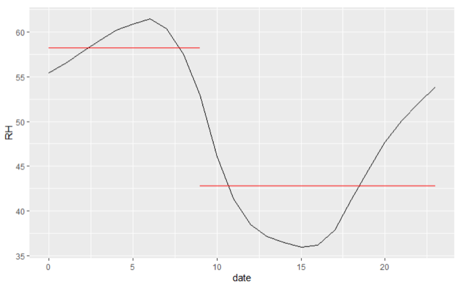


Figure 29: Changepoint in Rh

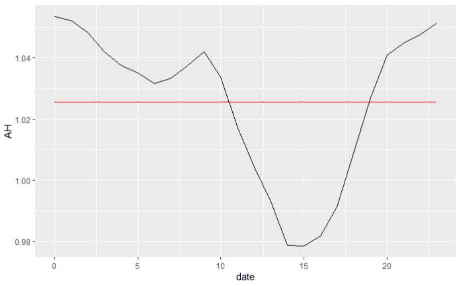


Figure 30: Changepoint in AH

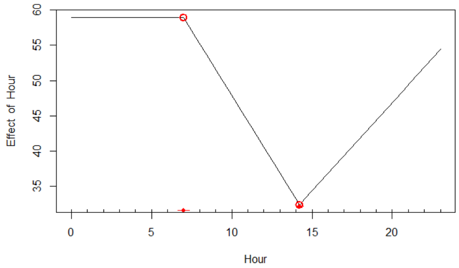


Figure 31: segmented in RH

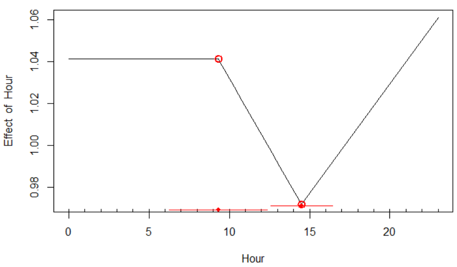


Figure 32: segmented in AH

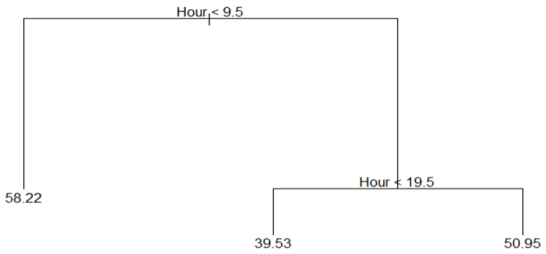


Figure 33: Tree for RH

and 7pm which is coherent with our previous results (Figure 33). However, the algorithm does not identify any change point in AH.

bcp

Finally, for the Bayesian approach, the lower posterior probability plot shows that at the locations 10, the probability of a change is very high for RH (Figure 34). We can get the exact locations where probabilities are high (e.g.> 70%). The probability of a change point at 10am equal to 0.74. However for AH, none of the hour have a probability of changing point higher than 70% (Figure 35).

Overall, three out of the four methods identify a change point around 10am

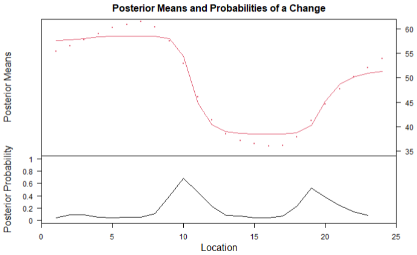


Figure 34: Bcp in RH

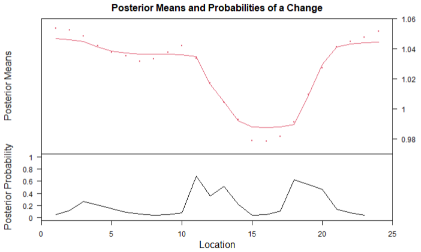


Figure 35: Bcp in AH

in RH which is coherent with the boxplot of RH. However, three of the methods suggests no change point for AH which also seems to be coherent when looking at the boxplot of AH. From the histogram, we can observe a severe drop around 2pm which is identified by the segmented method.

7 Summary and Conclusion

In the final section of this report we provide an overall view of the major conclusions that we have been able to draw from analysing this data set:

1. The multivariate regression model provided a decent fit to our data with our 10 predictors being able to explain most of the variance in our two outcome variables.

2. The residuals from our multivariate regression model were skewed to the right which suggested that the estimates obtained were biased. We conducted a simulation study to better understand this phenomenon and were able to conclude that the estimates from our multivariate regression model were robust even though we had skewed residuals.

3. The principal component analysis (PCA) of our data set helped us conclude that about 90% of the variance in our data set can be explained using just three principal components. Using these principal components as predictors in a PCR model we were able to get a competitive RMSE on the test set as compared to the multivariate regression model.

4. The principal component method and the maximum likelihood method agreed on the number of factors grouping the pollutants, the temperature and the nitrogen into three factors for the explanatory variables.

5. RH and AH have monthly seasonality and are also changing as per the time of the day, which could be because of the varying number of automobiles (emitting air pollutants). We observe a change point in RH around September and every day around 9am. AH peaks in the Fall.