The Analaysis of PM10 Suspended Partculate Air Pollution Using Several Robust Technique: Case Study in An Industrial Area, Penang, Malaysia

Habshah Midi*
Dept. Mathematics, Faculty of Science, UPM, Serdang, Malaysia - e-mail address

Hassan S. Uraibi
Laboratory of Computational Statistics and Operations Research, INSPEM ,UPM, Serdang, Malaysia
Dept. of Statistics, College of Administration and Economics, University of Al-Qadisiyah -IRAQ,
hssn.sami1@gmail.com

Abstract

The delegate of monitoring stations of the DOE, Malaysia; observed the air quality in atmosphere of Seberang Prai, Pulau Pinang during (2004-2007). The data set was depended on PM10 concentration as a response variable and seven independent covariates distributed into four pollutants and three meteorological. This paper describes a procedure for extracting a small set of potential covariates to explain response variable such as the PM10 of air quality. For better the future predictive ability and consistence selection, the set of explanatory covariates need to be reducing as possible as to the exact observed covariates, and the dependencies among these covariates are need to be low. In order to achieve these goals, all possible regression, combined with concentration algorithms to reduce the number of selected potential covariates to a necessary minimum, is developed to be robust against all type of outliers. In addition this robust cleaning all subset regression is applied to obtain responsible factors describing PM10 concentration level in air quality research.
Keywords: All subsets regression;Robust cleaning; DGK; Median ball ; RFCH.

1 Introduction

This search focuses on analysis and modeling the data of air quality status in Penang which located in the industrial area of Malaysia. Its well known that air quality status is describe and disseminated according to the Air Pollution Index (API). This index is a composite reflection of overall air quality based on five pollutants: sulfur dioxide (SO2), nitrogen dioxide (NO2), carbon monoxide (CO), ozone (O3), and suspended particulates matter of less than 10 microns in size (PM10). The first four indexes are reported in parts per million by volume (ppmv), but the PM10 is reported in micrograms per cubic meter of Air (ug/m3). These pollutants are mainly affecting the respiratory system and ecosystem. According to the report Department of Environment, Malaysian (DOE) the most prevalent pollutants records are PM10 and Ozone (O3).
The delegate of monitoring stations of DOE observed (1354) cases in atmosphere of Seberang Prai, Pulau Pinang during (2004-2007) was depended on depend PM10 concentration and seven independent variables, four pollutants and three meteorological (wind speed, temperature and relative humidity). So, we consider the sample size is large and there is possibility to use cleaning dataset techniques to obtain high breakdown and fast consistent estimator.
PM10 is the term given to the tiny particles of solid or semi-solid material found in the atmosphere which is a mixture of materials like smoke, soot, dust, salts, acids and metals. According to DOE, Malaysia, PM10 are emitted from heavy traffics, industries and open burning activities. Increase the concentration of PM10 in the atmosphere could cause severe effects health impacts to the human, particularly among the infants and elderly. People with respiratory illnesses such as asthma, nose and throat irritations, and allergies might cause premature mortality (for more details see: Fellenberg (2000), Godsh(2004),Tam et. al.(2004), and Baccini et. al. (2011)). Ul-Saufie et. al. (2012a) mentioned that Sedek et. al. (2006) found PM10 causes negative impacts on the growth and productivity of small and short cycle plant species such as vegetables. Thus, many researchers are focusing their studies in predicting PM10 concentration so that necessary preventative measures can be conducted.

Recently and fortunately predicting PM10 concentration give the attention of researchers in the statistical literature. The trials for developing models to predict PM10 concentrations due to the fact that statistical modeling could provide good insights in predicting future air pollutions index. Corani (2005) tried to predict the Ozone (O3) and PM10 using computational models which are feed forward neural networks (FFNNs), pruned neural network (PNNs) and lazy learning (LL). On the other hand, Pires et al. (2008) proposed 5 linear models which are multiple linear regression, principal component regression, independent component regression, quantile regression as well as partial least squares regression to predict the daily PM10 concentration.

Several approaches have been done to predict the PM10 concentrations based on the determination of the best model. The effort to find the best model is continued until in recent times where Ul-Saufie et al. (2012b) proposed a new method that combined both the regression models and back propagation models with principal component analysis (PCA). Saithanu et al. (2014) proposed using multiple linear regression with best subset and stepwise methods to predict PM10 concentration in Chonburi , Thailand.

The most commonly used method; linear models may not be the best model to use with the real data when there is outlier because it could violate the normality assumptions. Therefore, Ul-Saufie et al. (2012b) used modern robust statistical procedures that could be remedy the problem of presence of outliers and suggested robust regression model for prediction PM10 concentration in industrial area.

Variable selection is very important tool in the statistical inference. The goals of variable selection are to satisfy sensitivity (Richness, Adequacy), specificity (Parsimony, Sparsity), future predictive ability and selection consistency which we are willing one of models is true. It well known that many variable selection methods are relying on estimating the future predictive ability of each model.

Minimizing the criterion of $n^{-1}RSS(\beta)$, the average squared predictors error which is obtained from fitted model of interest on the observed sample, would be equivalent to maximizing coefficient of determinant, and always select the full model. Since the coefficients and models are being estimated and tested on the same data respectively. This case will lead to inflate assessment of the reduction in prediction error available from each predictor.

Mallows C_p (1973), is a powerful technique for model selection in linear regression. Mallows proposed adding an appropriate bias correction terms$2n^{-1}\sigma^2$ to n^{-1} RSS which can be defined as a linear function of Cp statistics as follows:

$$C_p = \frac{RSS_p}{\hat{\sigma^2}} - n + 2p \tag{1}$$

where RSS_p is the residual sum of squares for sub model p ,n is the number of observations, $\hat{\sigma^2}$ is an estimate of the error variance σ^2 which is usually counted in the full model. Since the low C_p indicates to a good future predicting ability where the $\hat{\sigma^2}$ is a good approximation for σ^2 of full model under linear regression assumptions. Subsets with smal C_p values have a small total mean squared error, and when C_p value is near$p$, the bias of the regression model is small. The shortcoming of this criterion, it is not clear whether is appropriate substitute to or any case can make the bias correction is not quite large enough will be tendency to overfit problem.

The Akaike Information Criterion (AIC) [1974] is proportional to the C_p statistic. The AIC is scaled in sum of squares unit. We search for models that have small values of AIC where the criterion is given by

$$AIC = -2\ell(\hat{\beta}_p) + 2p \tag{2}$$

A min AIC strategy is used for selecting among two or more competing models. In a general sense, the model for which AIC is smallest represents the best approximation to the true model. That it is the model with the smallest expected loss of information when MLEs replace true parametric values in the model. The AIC may be very poor where with small sample size which considered ($n < 40p$) by Burnham et $al.$ (2004), therefore, a small-sample correction on the penalty term could lead to the AIC_c statistic which proposed by Sugiura (1978) and Hurvich et al (1989). The AIC_c statistic is given by,

$$AIC_c = AIC + -\frac{2p(p+1)}{n-p-1} \tag{3}$$

For more information about AIC [see: Burnham et *al*. (2002, 2004), Hastie et*al*. (2001), and Kuha (2004)]. Bayesian Information Criterion (BIC) was introduced by Schwarz (1978) is driven from Bayesian theory to be by far the simplest to use. This criterion set prior probability for each possible model simultaneously set prior distribution and an independent prior probability for the coefficients of being nonzero in each model. Setting all prior probabilities are equal, and assuming that only one model along with its associated priors, is appropriate, Bayes theorem then can maximize the posterior probability of alternative models, given the data. Schwarz (1978) and Kashyap (1982) suggest criteria derived by taking a Taylor expansion of the log posterior probabilities of these models. These facts lead to the BIC which is,

$$BIC = -2\ell(\hat{\beta}_p) + log(n) + (p+2) \tag{4}$$

BIC tends to under-fit but it is consistent while, AIC and C tend to over-fit and inconsistent respectively. Robust selection criteria to compare a set of models has been received much attention in robustness literature by introduce robust versions of some criterions, for more details see e.g. Ronchetti (1997), Ronchetti, Field, and Blanchard (1997), Muller and Welsh (2005), Maronna et al. (2006), Salibian-Barrera and Van Aelst (2008), Claeskens and Hjort (2008), Heritier et al.(2009) and Tharmaratnam and Claeskens (2011).
Unfortunately, classical variable selections no long resist selecting the correct model in presence of outliers or other contamination. To overcome this problem, robust procedures are recommended. Robust selection criteria to compare a set of models has been received much attention in robustness literature by introduce robust versions of some criterions, for more details see e.g. Ronchetti (1997), Ronchetti, Field, and Blanchard (1997), Muller and Welsh (2005), Maronna et al. (2006), Salibian-Barrera and Van Aelst (2008), Claeskens and Hjort (2008), Heritier et al.(2009) and Tharmaratnam and Claeskens (2011).
Olive and Hawkins (2010) suggested ($RFCH$) that are used standard method for reweighting Fast Consistent High breakdown estimator (FCH), and gives easily computed $\sqrt{n}$ consistent outlier resistant estimator that can be used for inference. FCH estimator based on two attractors,DGK (Devlin, Gnanadesikan and Kettenring, 1981) and median ball Olive (2008) with some kind of location criterion. $RFCH$ estimator is a resistant to multivariate outliers estimator.

The DGK,MB and$RFCH$ were fast consistent, high breakdown or fast consistent and high breakdown estimators respectively. Each estimator built with what is called concentration algorithm which is convergence after five steps. The target of these algorithms obtains robust location and scatter matrix. The general framework DGK , MB and $RFCH$ algorithms split the dataset into clean and contaminate dataset. The main procedure of DGK and MB are that the algorithms should be find out the values of mahalanobis distance which are less than some cutoff point in each step, then estimate the robust location and scatter matrix. The last step repeats this procedure at least five times until convergence. The FCH estimator used DGK or MB as attractor with some criterion. $RFCH$ estimator is the reweighted version of FCH.
This article discussed robust linear regression variable selection for predicting PM10 concentration in Penang which is the most densed populated states in Malaysia with 1490 persons per square kilometre [12].
We consider the classical variable selection (subset selection) in linear regression as a main part in our proposed algorithm, that because selection predictors in linear regression allows simplified discussion for the most methods of interest. Variable selection is freer than model selection and more extensions and extensions to more general settings are often straightforward. (John Dziak, 2007).
We suggest modifying $DGK(MDGK)$ and $MB(MMB)$ algorithms to be fastest convergences than the last ones, then call the concentrated dataset in the last steps of $MDGK$,MMB,$MRFCH$ with classical variable selection methods, then incorporate the cleaning data of DGK, MB and$RFCH$ with all possible subset based on C_p Mallows, AIC, and BIC respectively. This approach will therefore be called cleaning variable selection and can be considered a trade-off between quality of data and model interpretability.
Our proposed algorithm aims to find a robust variable selection for the response PM10 that include low potential covariates and more accurate in order to achieve high interpretability. The rest of this paper is organized as follows. In Sect. 2, we will describe robust and fast consistent variable selection algorithms in more detail. Section 3 analyses the air quality dataset of Penang, Malaysia, 4 outlines how our proposed algorithm can be applied to obtain a small subset of explanatory variables determining PM10, and a simulation study is performed in Sect. 4. The final Sect. 5 concludes.

2 Description of Robust and Fast consisntent Variable selectin

Let a multivariate location and scatter model is a joint distribution of the ith case of $P \times 1$ random vector that completely specified by a $P \times 1$ population location vector μ and a $P \times P$ symetric positive definite population scatter matrix Σ. Assume that n cases are collected in a $n \times P$ matrix x, such that the $x_1^T, x_2^T, , x_n^T$ are independent. Consider the classical normal regression model $Y = x\beta + \varepsilon$, where Y is the vector of response variable, β is the vector of regression parameter and $\varepsilon \sim N(0, \sigma^2 I_n)$. The algorithm of robust and fast consistent variable selection consists of three main stages which can explain as follows:

Stage 1. Cleaning dataset.

The assumptions of cleaning algorithms assumed that the normality assumption of linear regression is violated by outliers or other contaminations, the sample size should be large, moderate multicollinearity between two or more than two covariates and independent residuals. Suppose matrix X is combined between the response vector Y and the covariates matrix x. We modified DGK, MB and RFCH would denote as $MDGK$, MMB and $MRFCH$ respectively to be suitable for cleaning X, and can be summarized as follows.

Algorithm 1: MDGK

The first step begins with classical estimator $(\bar{x}, cov)$ as initial or starts $(T_{(0,Start)}, C_{(0,Start)})$ to find the mahalanobis distance.

$$D_{(0,MDGK)} = \sqrt{(X - T_{0,start})^t (C_{0,start})^{-1} (X - T_{0,start})} \tag{5}$$

then reordering the observations of full dataset according to their mahalanobis distances, and find the halfset that include only the observation which have mahalanobis distance less than the median of whole mahalanobis distances,

$$Med_{(0,MDGK)} = Median(D_{(0,MDGK)}) \tag{6}$$

$$\widetilde{X}_{1,MDGK} = \{X_{ij} : D_{(0,MDGK)} \leq Med_{0,MDGK}\} \quad i = 1, 2, ..., p; \quad j = 1, 2, ..., m \tag{7}$$

Let $C_{(0,MDGK)} = C_{(0,Start)}$, again recalculate the average and variance-covariance estimators of $\widetilde{X}_{(1,MDGK)}$ halfset to get the first attractor $T_{(1,MDGK)}, C_{(1,MDGK)}$. If the diagonal elements of $C_{(1,MDGK)} = C_{(0,Start)}$ stop the algorithm, if not repeat the procedures until convergence to get the final attractor $T_{(K,MDGK)}, C_{(K,MDGK)}$ and $\widetilde{X}_{(K,MDGK)}$ where K the convergence step.

Algorithm 2. MMB

Suppose the initial variance- covariance matrix $C_{(0,Start)} = diag(p)$, and the Med is the median vector of matrix X , then the mahalanobis distances as follows,

$$D_{0,MMB} = \sqrt{(X - Med)^t (C_{0,start})^{-1} (X - Med)} \tag{8}$$

Now, let the location criterion cutoff is the median of $D_{(0,MB)}$, that denoted as $luct$,

$$luct = Med_{(0,MMB)} = Median(D_{0,MMB}) \tag{9}$$

when $luct \neq 0.5$, the cutoff point should be the quantile of $D_{0,MB}$ with probability equals 0.5.
The concentration for cleaning X need to find the half dataset with only non outlying observations which have mahalanobis distance less than or equals the median of it,

$$\widetilde{X}_{1,MMB} = \{X_{ij} : D_{0,MMB} \leq Med_{0,MMB}\} \quad i = 1, 2, ..., p; \quad j = 1, 2, ..., m \tag{10}$$

Now get start with $T_{1,MBA}$ is the average of $\widetilde{X}_{1,MB}$ and $C_{1,MBA}$ is variance-covariance matrix of it. For more concentrations we can recalculate the mahalanobis distances and repeat the procedure until convergence to get the final attractor $(T_{K,MMB}, C_{5,MB})$ and $\widetilde{X}_{K,MB}$ where K the convergence step.

Algorithm 3. Modified Reweighted Fast and Consistent High Breakdown $(MRFCH)$

Olive et al. (2010) developed the idea of MB by adding the location criterion or cutoff point to select the

attractor, and proposed what is so called Modified Fast Consistent and High breakdown (FCH) estimator. We modified FCH based on the final attractors od $MDGK$ and MMB. If the Euclidian distance between the $MDGK$ location $T_{K,MDGK}$ and $MED(X)$ less than or equals to $MED(D_i(MED(X), I_p))$ the MFCH estimator uses only the with tattractorhe smallest determinant as follows,

$$T_{MFCH} = \begin{cases} T_{K,MDGK} & if \quad \sqrt{|C_{K,MDGK}|} < \sqrt{|C_{K,MMB}|} \\ T_{K,MMB} & if \qquad\qquad Otherwise \end{cases}$$

and the scale as follow,

$$C_{MFCH} = \begin{cases} \frac{MED(D_i^2(T_{K,MDGK}, C_{K,MDGK}))}{\chi^2_{(p.0.5)}} \times C_{K,MDGK} & if \quad \sqrt{|C_{K,MDGK}|} < \sqrt{|C_{K,MMB}|} \\ \frac{MED(D_i^2(T_{K,MMB}, C_{K,MMB}))}{\chi^2_{(p.0.5)}} \times C_{K,MMB} & if \qquad\qquad Otherwise \end{cases}$$

where $\chi^2_{(p.0.5)}$ is the 50th percentile of a chi-square distribution with p degrees of freedom.

Reweighted $MFCH$ attractors by isolate the observation with $D_i^2(T_{MFCH}, C_{MFCH}) \leq \chi^2_{(p,0.975)}$, the using the classical estimator to obtain $T_{1,MFCH}$ and $C_{1,MFCH}$ from,

$$\widetilde{X}_{1,MFCH} = \{X_{ij} : D_i^2(T_{MFCH}, C_{MFCH}) \leq \chi^2_{(p,0.975)}\} \quad i = 1, 2, ..., p; \quad j = 1, 2, ..., m \tag{11}$$

The new cutoff point is that $\frac{MED[D_i^2(T_{1,MFCH}, C_{1,MFCH})]}{\chi^2_{(p,0.5)}}$ and the new variance covariance matrix is,

$$C_{2,MFCH} = \frac{MED[D_i^2(T_{1,MFCH}, C_{1,MFCH})]}{\chi^2_{(p,0.5)}} \times C_{1,MFCH} \tag{12}$$

Reweighted the estimators again by repeat the equation . with new cutoff point to get the final attractors (T_MRFCH, C_MRFCH) and $\widetilde{X}_MRFCH$ which is consists of from the only the concentrate observation of response and independent variables.

Stage 2. Best Subset Regression Selection
Although it takes longer to run, the all subset regression guarantees including all potential observed covariates at least in one subset. The classical AIC, and BIC criterions topically have the ability to determine the best subset. When the assumptions of cleaning are met in stage one, the all subset methods are usage to select the best subset. The three refined datasets in stage one are constructed based on $MDGK, MMB$, and$MRFCH$ algorithms respectively. Actually AIC, and BIC criterions will run for each refined dataset to select the best subset.

Stage 3. Adjustment the Best Subset Regression Selection This stage suggests for adjustment the best subsets that were selected by AIC criterions in stage two. Fitting the regression model for each subset based on the observed dataset which are selected in stage one, the select only the coefficients that associates with p-vlue less than $^* = 0.05/d$ where d is the number of all candidate covariates. We proposed this procedure to overcome the problem of over-fit in AIC selection, and we expect, it will regular the performance of it to be consistence.

3. Air quality data
A real data of the annual hourly observations for $PM10$ in Seberang Prai, Pulau Pinang from January 2004 to December 2007 was taken from Department of Environment. This hourly data were transformed A real data of the annual hourly observations for PM10 in Seberang Prai, Pulau Pinang from January 2004 to December 2007 was taken from Department of Environment. This hourly data were transformed into daily data by taking the average and median PM10 concentration level for each day. Figure 1, both histogram and qq-plot show that residuals are contaminated by with heavy tails mixture distribution. Since some points in

qq-plot is not in the straight line it indicates that these points are not normal. Thus, we suspect that there are outliers in the average data air quality dataset of Seberang Prai, Pulau Pinang atmosphere. Figure 1, D indicate to existence some high leverage points in the the relationship between PM10 and O3. All subfigures except D are no clear whether, including leverage point unless using the diagnostic method. TTo

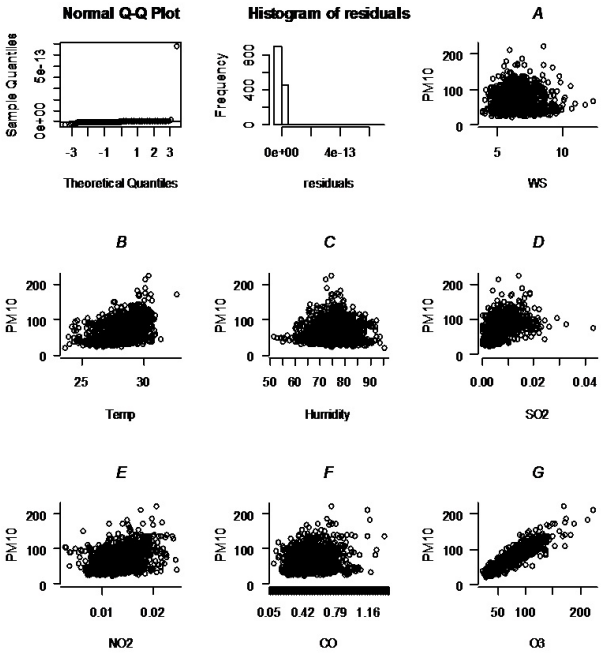

Figure 1: QQ-Plot, Histogram of Residuals and plot $PM10$ vs each the component of air quality data

cut removes all doubt that there are outliers violated the normality assumption. We note in Figure 2-(b) the robust mahalanobis distance is identified some leverage points, while the classical one fail to identify all of them in Figure 2-(a). The $DRGP$ method (Habshah et al, 2009) is employed to average air quality data to identify the high leverage points. DRGP detected 112 high leverage points for average daily data. The PM10 concentration data is not normal as they contain leverage points. Thus using the Ordinary Least Squares (OLS) method in estimating the parameters of the multiple Linear Regression will led to misleading conclusion. Cleaning algorithm base$MDGK$ and MMB algorithms identified only 677 clean observations. Figures 3 and 4 the normal $Q-Q$ plots observe that the all points are enough close to the line; close enough to say that these 677 clean observations coming from normal distribution, and the histogram in both figures show the residuals are normally distributed. We note that the subfigures A to G in Figure 3 and 4 are seem the same, we think $MDGK$ and MMB algorithms diagnose the 677 clean observation which represent around 50% sample size.

On the other hand, $MRFCH$ algorithm is chosen 1167 clean observation. We note that the $MRFCH$ algorithm considered only 187 observation as outliers and added 500 observation which excluded by other algorithms to cleaning observations. We observe in Figure 5 normal $Q-Q$ plot that, this augment in sample size of cleaning dataset results in the points be quit close to the line of normal distribution and histogram of residuals too. The subfigures A to G in Figure 5 differed with previous subfigures (Figure 3 and 4) as a result of the increasing in the sample size of clean air quality dataset. Collection the air quality data was taken in account the time factor, which requires checking the autocorrelation problem, and whether there is multicollinearity problem. The statistics of Durbin Watson ($D.W$) test of autocorrelation and VIF test of multicollinearity for three cleaning algorithms are present in table 1 below. The inferred from the values shown in table 1, that the three clean datasets were collected with independent variables and independent

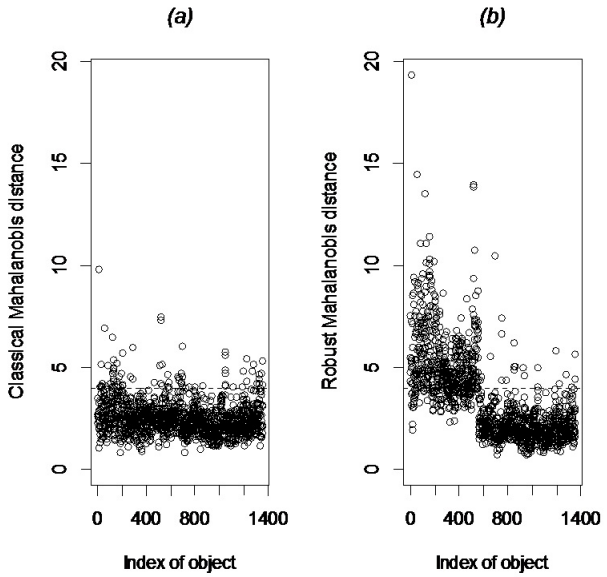

Figure 2: Classical and Robust mahalanobis distance for Air quality dataset.

residuals.

Table 1: D.W and VIF for cleaning dataset based on MDGK, MMB, MRFCH algorithms.

	$D.W$	p -value	VIF
Original	2.132	0.992	2.104
MDGK	2.040	0.619	9.245
MMB	2.075	0.567	8.198
MRFCH	2.075	0.683	9.477

Next, we would like to further investigate the important variables that influence the PM10 by employing our proposed method and other existing methods. Our proposed methodology is carried out with the following setting. The data set consists of 1354 observations which include the PM10 as the response variable and seven independent variables already mentioned. We divide the data into training and test sets whereby 70% and 30% of the data is randomly chosen as training and test sets. This process is repeated for 500 times. The$MRFCH$ was used to clean the training and test set data. The classical all subset selection selects the best model of training set based on the AIC and BIC criterions. Subsequently, the significant variables of the best model of the training set are determined by observing the coefficients whose p-values are less than 0.007. It is important mentioning that only stage 1 of our proposed method was applied to the test sets data. For cross validation, the coefficients of each training model are used to predict the response (PM10) with the test set data. The residuals and the root mean square errors ($RMSE$) were then computed. Table 2 and Table 3 exhibit the results of our proposed methods and existing methods before and after adjustment of the significant levels. Each table shows the average $RMSE$ for 500 test sets data and the percentage of each variable being selected in a model over 500 training sets data. The threshold values were computed based on the central point of the 50th and 75th percentiles of the percentage of each variable being selected for each method. Following Meinshausen and Buhlmann (2009), threshold is defined as:

$$\pi_{th} = \sqrt{EV \times p \times (2 \times \lambda - 1)}/p \tag{13}$$

where EV the expected number of falsely variables selected. p is the number of coefficients, and λ=0.80 .

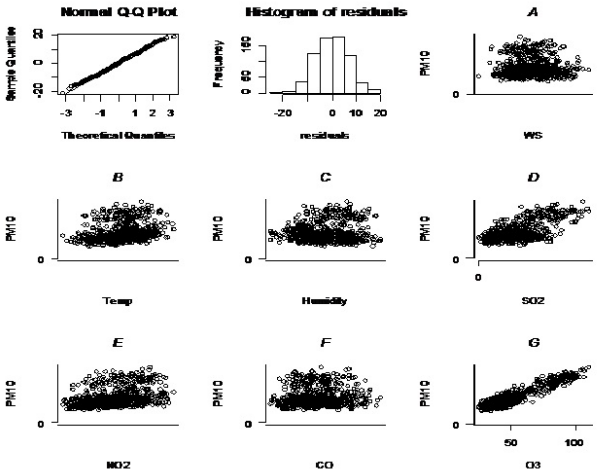

Figure 3: QQ-Plot, Histogram of Residuals and plot PM10 vs. each the components of air quality for 677 observations are chosen by MDGK algorithm.

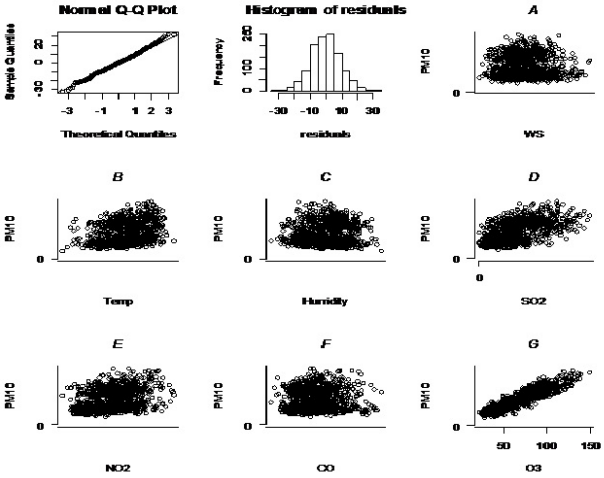

Figure 4: QQ-Plot, Histogram of Residuals and plot PM10 vs. each the components of air quality for 677 observations are chosen by MMB algorithm.

The threshold values in Tables (2-3) is calculated as follows:

$$\pi_{th} = \sqrt{3 \times 7 \times (2 \times 080 - 1)}/7 = 57.14$$

The potential candidate variable is the one that has percentage of being selected in a model exceeds the threshold value. The best method is the one that has the lowest average of RMSE.

The results of Tables 2-3 signify that the $RMSE$ of our proposed methods based on both AIC and BIC (before and after adjustments of significant levels) is the smallest compared to other methods (all subsets AIC and BIC). This suggests that our proposed method correctly identify the potential variables namely $NO2, CO$, and $O3$ to be included in the final model. The all subset- AIC and all subset- BIC methods selects four potential variables. It is interesting to observe that all methods only select pollutants variables

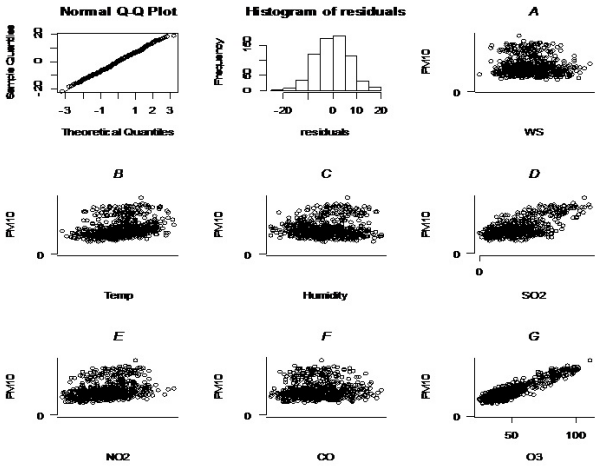

Figure 5: QQ-Plot, Histogram of Residuals and plot PM10 vs each the components of air quality for 1167 observations are chosen by MRFCH algorithm. .

in the final model. The meteorological variables are not selected in the final model perhaps they do not have much influenced on the PM10 variable. The results also clearly indicate that $O3$ have a very strong correlation with the PM10. The percentage of this variable to be included in the final model is 100%. It can be observed that theAIC tend to over fitting model and BIC tends to under fitting model. For example, for variable WS, the percentage of selecting this model is 6.40 but it only 0.20 for BIC criterion.

The results of the model validation suggests that NO2, CO, and O3 are finally chosen to be included in the final model. Once the model has been validated, the entire data set of the original data is used for estimating the regression model. The Ordinary Least Squares (OLS) method is often used to estimate the parameters of a model. The PM10 model (OLS) is given by the following equation:

$$PM_{10} = -0.0579\ NO_2 - 0.0548\ CO + 0.9606\ O_3 \tag{14}$$

The standard errors of each estimates, $NO2, Co$ and $O3$ are 0.013, 0.018, 0.012 respectively.

It is now evident that outliers have an adverse effect on the OLS estimates. As an alternative we suggest to use robust method to estimate the parameters of the model. Here, we suggest to employ a high efficient and high breakdown point MM estimator. The PM10 model (MM) is given by the following equation:

$$PM_{10} = -0.066\ NO_2 - 0.045\ CO + 0.9606\ O_3 \tag{15}$$

The standard errors of each estimates, NO2, Co and O3 are 0.012, 0.011. 0.012 respectively.

We wish to compare the two models based on their standard errors. The final model selection is based on the standard errors of each estimate. Since the standard errors of the parameter estimates of the second model are smaller than the first model, Model 2 is ultimately chosen as the final model.
Both tables (2,3) are shown the lowest $RMSE$ of all subset the ones running with cleaning algorithm $MRFCH$. The method which based onBIC selects only three pollutants ($NO2, COandO3$) before and after adjustment. On other hand robust lars select $SO2, CO$ and $O3$ but the $RMSE$ is 0.388. We have noticed that the meteorological variables have not been a big opportunity to be displayed on the model chosen. Perhaps this is due to their inability to influence PM10 variable, and therefore presence or absence of outliers with metrological variables does not make them competitive with the rest of variables (pollutants). It is very clear in both tables ozone variable is associated with a very strong correlation with the $PM10$, where the presence of outliers or not, $O3$ have a chance of being in the model chosen by 100%. Unlike

Table 2: Selected variables, the average rmse , the percentage of selecting variable of four all subsets methods before adjustment the significant level.

	$RMSE$	WS	$Temp$	$Humidity$	SO_2	NO_2	CO	O_3	Selected $Var.$
$Allsubset.aic$	0.388	6.40	6.60	20.4	69.8	99.8	100	100	4,5,6,7
$Allsubset.bic$	0.388	0. 20	0.20	1.20	17.4	87.2	94.8	100	5,6,7
$MRFCH.aic$	0.320	36.0	30.6	7.00	71.0	100	9.88	100	4,5,6,7
$MRFCH.bic$	0.320	2.00	3.60	2.00	28.6	99.4	72.6	100	5,6,7

previous methods Lars nominated $SO2$ variable to be in model chosen, rather than the $NO2$ variable. This nomination is unfortunate according to the percentage of the presence for $SO2$ and $NO2$ in the final model which is chosen by rest of methods. We observed that the presence of$NO2$ in the best model, have more stability and stable than $SO2$ in the previous methods whether robust or non-robust of it. Finally, the final model results from the procedure with lowest $RMSE$ and high percentage of selecting variables contains the predictors $NO2, CO$, and $O3$ with the following fitting model.

$$PM10 = -0185\ NO_2 - 0.0476\ CO + 0.7949\ O_3 \tag{16}$$

Table 3: Selected variables, the average RMSE , the percentage of selecting variable of four all subsets methods after adjustment the significant level to 0.007.

	$RMSE$	WS	$Temp$	$Humidity$	SO_2	NO_2	CO	O_3	Selected $Var.$
$Allsubset.aic$	0.388	4.00	7.60	20.8	66.92	99.6	100	100	4,5,6,7
$Allsubset.bic$	0.388	4.00	7.60	20.8	66.92	99.6	100	100	4,5,6,7
$MRFCH.aic$	0.322	0.00	00.0	1.00	15.6	92.8	95.2	100	4,5,6,7
$MRFCH.bic$	0.322	1.20	0.26	0.00	17.4	92.8	95.2	100	5,6,7

3 Simulation

A design matrix coming from a centered multivariate normal distribution with covariance structure is considered $cov(X_j; X_K)^{|j-k|}$ where $\rho = 0.5$, $j = 1, 2, ..., 10$ and $k = 1, 2, ..., 10$. The response variable Y is generated using the following equation,

$$Y = 7X_1 + 6X_3 + 5X_4 + 7X_6 + 7X_9 + 0X_D + \epsilon \tag{17}$$

where D=2,5,7,8,10

egarding contamination, we follow Claudio et al., 2010 cases, whereε denotes the fraction of outliers in the data, for no contamination case the residual is standard norma $\epsilon \sim N(0, 1)$. The second contamination case is that contaminated the residuals by ε symmetric outliers with the Slash distribution where ε=0.10;$\epsilon \sim (1 - \varepsilon)N(0, 1) + \varepsilon N(0, 1)/u(0, 1)$. The last case contaminates Y by ε asymmetric outliers $\epsilon \sim (1 - \varepsilon)N(0, 1) + \varepsilon N(20, 1)$, but contaminated observations contain outliers in $X_1, X_2, , X_10$ coming from$N(50, 1)$. For each case we generated 500 independent simulated dataset. We spilt each of these dataset randomly into training n^{tr} and test n^{ts} sets, and repeat this procedure 50 times. Each training set n^{tr} , we use the usual all subsets method based on BIC ($Allsubset.bic$) as implemented in R package leaps. The all subsets method based on AIC($Allsubset.aic$), we used the same package by subtract the constant $log(n)p + 2p$ from BIC values. The minimum value of AIC or BIC certainly will determine the significant predictors. Fit the model of choice with and adjusted p-value of the model coefficients. The variable which poses p-value less than 0.005 is chosen in the final model. These significant variables are recorded to know the

number of times of selected by each method separately, and then we take the average over 50 training sets. Finally the relative average are computed to determine the opportunity of each predictor remind in the final model. Again Fit the model with the most significant variables, and take the advantage of its coefficients for error prediction with test setn^{tr} . The average of root of mean square error for 50 test sets is put forward estimate the average of average of mean square error over all 500 independent simulated dataset. Actually, $MRFCH$ algorithm clean up training set n^{tr} from outliers and use the remain observations as new set. Let it is denoted as n_{tr}^{RFCH} which is less than n^{tr} . The $MRFCH.aic$ and $MRFCH.bic$ follow the same procedure mentioned above with new set n_{tr}^{RFCH}. The cleaning process include the test set n^{tr}.

For no contamination case the results are present in table (4) which is shown that all methods performed very close to each other. Table (5) shows the results of our simulation with 10% symmetric vertical outliers. Deciding the best method is that ones minimize $RMSE$ and select the correct variables in the final model. All methods select the correct variables, but our proposed methods minimize $RMSE$ to be less than others. On other hand where the leverage points appear in the dataset together with vertical outliers result in over-fit problem to the classical methods (see table 6). From the results in table (6) we infer that MRFCH.aic and MRFCH.bic performance better than the classical ones and consistent. Since we observe selecting the correct variable is stable and the opportunity to choose the noise is virtually nonexistent.

Table 4: Selected variables, the average $RMSE$, the percentage of selecting variable of four all subsets methods after adjustment the significant level to 0.005 and for type one contamination (clean)

	$RMSE$	1	2	3	4	5	6	7	8	9	10	Selected $Var.$
Allsubset.aic	0.67	100	17.9	100	100	15.3	100	15.4	15.2	99.9	16.6	1,3,4,6,9
Allsubset.bic	0.67	100	1.79	99.9	99.7	1.65	99.9	1.51	1.47	99.9	1.71	1,3,4,6,9
$MRFCH.aic$	0.65	99.9	0.84	99.7	99.6	0.84	99.7	0.88	0.75	99.8	0.81	1,3,4,6,9
$MRFCH.bic$	0.65	99.9	0.84	99.7	99.6	0.84	99.7	0.88	0.75	99.8	0.81	1,3,4,6,9

Table 5: Selected variables, the average RMSE , the percentage of selecting variable of four all subsets methods after adjustment the significant level to 0.005 and for type one contamination (sym)

	$RMSE$	1	2	3	4	5	6	7	8	9	10	Selected $Var.$
Allsubset.aic	0.66	91.2	17.0	76.4	88.4	18.1	85.9	17.9	15.6	68.4	18.3	1,3,4,6,9
Allsubset.bic	0.66	87.5	3.70	63.8	84.3	4.21	78.7	3.28	3.02	51.9	3.70	1,3,4,6,9
$MRFCH.aic$	0.21	100	1.14	100	100	1.22	100	1.21	0.78	99.5	0.95	1,3,4,6,9
$MRFCH.bic$	0.21	100	1.14	100	100	1.22	100	1.21	0.78	99.5	0.95	1,3,4,6,9

Table 6: Selected variables, the average RMSE , the percentage of selecting variable of four all subsets methods after adjustment the significant level to 0.005 and for type two contamination (LP)

	$RMSE$	1	2	3	4	5	6	7	8	9	10	Selected $Var.$
Allsubset.aic	0.039	100	99.89	100	100	17.22	100	15.8	19.38	100	16.65	1,2,3,4,6,9
Allsubset.bic	0.039	100	97.72	100	100	1.72	100	2.08	2.79	100	2.66	1,2,3,4,6,9
$MRFCH.aic$	0.036	100	2.16	100	100	1.04	100	0.49	1.31	99.9	1.27	1,3,4,6,9
$MRFCH.bic$	0.036	100	2.16	100	100	1.04	100	0.49	1.31	99.9	1.27	1,3,4,6,9

4 Conclusion

The practical application motivated us to develop all subsets method for finding robust linear regression selection which is limited only the variables that describe the response. Our method is based on RFCH procedure of Olive et al. (2010). Rather than replacing original dataset by clean ones and applying the same all subsets method. Several methods for all subsets selection are available to date, but only a few proposals for robust all subsets selection. Our simulation studies suggest that our method is robust to the presence of vertical outliers, and leverage points in the data, and that in these cases it compares well with classical all subsets selection based on AIC and BIC respectively.
From the results presented in the tables noticed that the classical all subsets method chooses the correct variables in most cases, but $RMSE$ higher than the robust ones. Actually, this may be due to cross validation procedure. Split the data set into two different sets of training and test size by 70% and 30% respectively, probably make the training set unaffected by outliers. This case when the proportion of outliers in training set is less than 5%. In any case, our proposed method is more stable and efficient than the traditional all subset selection.

References

Agostinelli, C.,Salibian Barrera,M. (2010). Robust Model Selection with LARS Based on S-estimators. Proceedings of COMPSTAT'2010,pp 69-78.
D.J. Olive, (2004). A resistant estimator of multivariate location and dispersion, Computational Statistics and Data Analysis 46, pp. 99102.
D.J. Olive and D.M. Hawkins (2010). Robust Multivariate Location and Dispersion.
http://lagrange.math.siu.edu/Olive/pphbmld.pdf. December.
D.J. Olive (2008). Applied Robust Statistics, preprint available from (http://lagrange.math.siu.edu/Olive/run.pdf).
Ronchetti, E. (1985). Robust model selection in regression. Statistics and Probability Letters, 3: 2123.
Ronchetti, E., Field, C., and Blanchard, W. (1997). Robust linear model selection by cross-validation. Journal of the American Statistical Association, 92: 10171023.
Ronchetti, E. and Staudte, R. G. (1994). A robust version of Mallows . Journal ofthe American Statistical Association, 89: 550559.
S.J. Devlin, R. Gnanadesikan, and J.R. Kettenring, (1981). Robust estimation of dispersion matrices and principal components, Journal of the American Statistical Association 76 pp. 354362.