Journal of Data Science 13(2015), 663-692

THE COMPARISON OF PARTIAL LEAST SQUARES REGRESSION, PRINCIPAL COMPONENT REGRESSION AND RIDGE REGRESSION WITH MULTIPLE LINEAR REGRESSION FOR PREDICTING PM10 CONCENTRATION LEVEL BASED ON METEOROLOGICAL PARAMETERS

Esra Polat#1, Suleyman Gunay*2

 # Department of Statistics, Faculty of Science, Hacettepe University, 06800, Ankara, Turkey

[1]espolat@hacettepe.edu.tr

Abstract: Air pollution shows itself as a serious problem in big cities in Turkey, especially for winter seasons. Particulate atmospheric pollution in urban areas is considered to have significant impact on human health. Therefore, the ability to make accurate predictions of particulate ambient concentrations is important to improve public awareness and air quality management. Ambient PM10 (i.e particulate diameter less than 10um in size) pollution has negative impacts on human health and it is influenced by meteorological conditions. In this study, partial least squares regression, principal component regression, ridge regression and multiple linear regression methods are compared in modeling and predicting daily mean PM10 concentrations on the base of various meteorological parameters obtained for the city of Ankara, in Turkey. The analysed period is February 2007. The results show that while multiple linear regression and ridge regression yield somewhat better results for fitting to this dataset, principal component regression and partial least squares regression are better than both of them in terms of prediction of PM10 values for future datasets. In addition, partial least squares regression is the remarkable method in terms of predictive ability as it has a close performance with principal component regression even with less number of factors.

Keywords: air pollution, meteorological parameters, model fit, multicollinearity, multiple linear regresion, principal component regression, partial least squares regression, ridge regression, particulate matter (PM10), prediction.

1. Introduction

In several linear regression and prediction problems, the independent variables may be many and highly collinear. This phenomenon is called multicollinearity and it is known that in the case of multicollinearity the ordinary least squares (OLS) estimator for the regression coefficients or predictor based on these estimates may give very poor results. Therefore, several biased prediction methods have been developed to overcome multicollinearity problem such as ridge regression (RR), principal component regression (PCR) and partial least squares regression (PLSR). The main goal of biased methods is to decrease the mean squared error of prediction by introducing a reasonable amount of bias into the model. In most real systems, exact collinearity of variables in X is rather unusual, because of the presence of random experimental noise. Nevertheless, in systems producing nearly collinear data, the solution for regression coefficient is highly unstable, such that very small interferences in the original data (for example, because of noise or experimental error) cause the method to produce madly different results. In addition, the use of highly collinear variables in multiple linear regression (MLR) also increases the possibility of overfitting the model. Overfitting means that the model may fit the data very well, but fails when used to predict the properties of new samples (Martens and Naes, 1989; Naes et al., 2002).

PLSR is a method that extracts the latent variables (LVs), which serve as a new predictors and regresses the dependent variables on these new predictors. PLSR comprises of regression and classification tasks as well as dimension reduction techniques and modelling tools. Therefore, it could be applied as a discrimination tool and dimension reduction method similar to principal component analysis (PCA) (Rosipal and Krämer, 2006). PLSR may overcome the collinearity problem with fewer factors than PCR. In addition, it requires less computation than PCR. Meanwhile simulations tend to show that usually PLSR reaches its minimal mean square error (MSE) with a smaller number of factors than PCR (Helland, 1988).

Air pollutants may cause changes in atmospheric composition and chemistry, which results in global warming, ozone depletion, dry and wet deposition and unwanted effects to human, animal, plant and material. Emissions from mobile sources (i.e. motor vehicles) have been one of the major sources of air pollution in some big cities (Tasdemir et al., 2005). Air pollution is one of the important issues in urban and industrial areas. The adverse effects of air pollutants have become a common problem in environmental sciences because of the major environmental risk to health in many developed or developing countries. Many air pollution indicators affect human or environment health. Some of them are such as particle pollution (often referred to as particulate matter), ground-level ozone, carbon monoxide, sulfur oxides, nitrogen oxides and Pb are defined as criteria air pollutants by the U.S. Environmental Protection Agency (US EPA) based on human/environment health impacts and cause property damage (Ozdemir and Taner, 2014).

Particulate matters (PM) including the solids and liquids dispersed in air are originated from different sources. For Example, coarse particles (>10 um) are products of mechanical processes, grinding, spraying and wind erosion (Tasdemir et al., 2005). PM10 (i.e. particulate diameter less than 10um in size) is one of the major components of air pollution that threatens both our health and our environment. Of greatest concern to public health are the particles small enough to inhale into the deepest parts of the lung such as PM10. PM10 pollution consists of very small liquid and solid particles floating in the air. These particles are less than 10 microns in diameter - about 1/7th the thickness of a human hair. Due to consuming solid fuel for heating, the amount of PM10 is increasing during winter seasons (October-March) while decreasing during summer seasons (April-September) (Polat, 2009). The ability to accurately model and predict the ambient concentration of PM is essential for effective air quality management and policies development (Sayegh et al., 2014). There is an extensive literature on modelling PM10 concentrations for different cities of different countries using different models. Ian G. McKendry (2002) compared Multi-layer perceptron (MLP) artificial neural network (ANN) models with traditional MLR models for daily maximum and average particulate matter (PM10 and PM2.5) forecasting in Chilliwack (in the eastern Lower Fraser Valley), in Canada. He found that meteorological variables (precipitation, wind and temperature), persistence and co-pollutant data were useful PM predictors. He stated that “if MLP approaches are adopted for PM forecasting, training methods that improve extreme value prediction are recommended”. Archontoula Chaloulakou et al. (2003) evaluated based on a data inventory, in a fixed central site in Athens, Greece, ranging over a two-year period and using mainly meteorological variables (surface temperature, relative humidity, horizontal wind speed and wind direction) as inputs, neural network (NN) models and MLR models were developed. Siegfried Hörmann et al. (2005) investigated the influence of meteorological as well as anthropogenic factors on PM10 in Graz for the winter seasons 2002/03 (182 days) and 2003/04 (183 days) by using MLR. Moreover, they introduced a prediction model using current information and meteorological forecasts to predict the average concentration of PM10 for the next day. They found that PM10 concentration in Graz was highly influenced by three meteorological factors inversion, precipitation, wind and by human impacts like traffic, industry, households. Salimol Thomas and Robert B. Jacko (2007) developed a linear regression model and a NN model to forecast hourly PM2.5 and CO concentrations for Borman Expressway, which is a heavily traveled 16-mi segment of the Interstate 80/94 freeway through Northwestern Indiana, by using the year 2002 traffic, pollutant and meteorological data (wind speed, wind direction and temperature). The performance of these models were evaluated using the year 2003 dataset. A. Afzali et al. (2008) analyzed the correlation between meteorological parameters (temperature, wind speed, relative humidity and solar radiation) and PM10 concentrations by using MLR and ANN. This study presented the results of predicting ambient PM10 concentration and the influence of meteorological parameters based on the data sampled from 2008-2010 in an industrial area of PasirGudang, Johor. N. F. F. Md Yusof et al. (2008) investigated the relationship between the weather parameters such as wind speed, ambient temperature, relative humidity and the level of PM10 emission in Seberang Perai, Penang in different season (wet and dry season). Haze events had occurred in 2004. Therefore, hourly PM10 concentration data for that year were used for this analysis. The hourly PM10 datasets were then modeled using linear regression models to predict future PM10 concentration which incorporate with the weather parameters. The effects of haze event to the prediction were also investigated. The results of linear regression model indicated that PM10 concentration was affected by the weather parameters as well as haze event. M. Demuzere et al. (2009) studied the relations between meteorological actors and air quality on a local scale based on observations from four rural sites in the Netherlands and these relations were determined by a comprehensive correlation analysis and a MLR analysis in 2 modes, with and without air quality variables as predictors. I. Barmpadimos et al. (2011) investigated the trends of PM10 and the effect of meteorology for 13 air quality stations in Switzerland for the period between 1991 and 2008 by constructing Generalised Additive Models (GAMs) which incorporate non-parametric relationships between

PM10 and meteorological and time variables. These relationships were based on quality-checked meteorological and PM10 long-term observations including a wide spectrum of meteorological variables with good spatial and temporal coverage. A. Z. Ul-Saufie et al. (2011a) compared a MLR model and Feedforwad Backpropagation ANN for predicting PM10 concentration in Seberang Perai, Pulau Pinang. Pulau Pinang state is situated on the north-western coast of peninsula Malaysia. Annual hourly observations for PM10 from January 2004 to December 2007 were selected for predicting PM10 concentration level. The hourly observations were transformed into daily data by taking the average PM10 concentration level for each day. Since the data for September 2007 was missing, not included in the analysis. Relative humidity, wind speed, nitrogen dioxide (NO2), temperature, carbon monoxide (CO), sulphur dioxide (SO2), ozone (O3) and previous day PM10,t-1 were used as independent variables. Assessment of model performance indicated that NN could predict PM better than multiple regressions. However, models adequacy checked by various statistical methods showed that the developed multiple regression models could also be used for prediction of PM10. A. Z. Ul-Saufie et al. (2011b) aimed to to improve the predictive power of MLR models using principal components (PCs) as input for predicting PM10 concentration for the next day. The developed model was compared with MLR models. Annual hourly observations for PM10 in Seberang Prai from January 2004 to December 2007 were selected for predicting PM10 concentration level as in their previous study (2011a). As in their previous study (2011a), they selected the same independent variables except O3 to study the influence on PM10 concentration. MLR was used to predict the next day PM10 concentration using as predictors air pollutant (NO2, SO2, CO and PM10,t-1) and meteorological parameters. Two different approach were used, considering original data and PCs as inputs. The result showed that the usage of PCs as input provides more accurate results than original data because it reduced the number of inputs and therefore decreased the model complexity. Besides that, the use of PCs based models was considered more efficient, due to elimination of collinearity problem and reduction of the number of predictor variables. Assessment of model performance indicated that PCR can predict PM better than multiple regressions. However, models adequacy checked by various statistical methods showed that the developed multiple regression models can also be used for prediction of PM10. Voukantsis et al. (2011) also analyzed the PM2.5 and PM10 concentrations. However, they proposed a methodology consisting of specific computational intelligence methods, i.e. PCA and ANNs, in order to inter-compare air quality and meteorological data, and to forecast the concentration levels for environmental parameters of interest (air pollutants). They demonstrated these methods to data monitored in the urban areas of Thessaloniki and Helsinki in Greece and Finland, respectively. The dataset used in this study corresponded to the time period 2001–2003. Doreena Dominick et al. (2012) aimed to determine the influence of meteorological parameters (ambient temperature, relative humidity and wind speed) based on a daily average computation of air pollutants PM10 at three selected stations in Malaysia, namely Shah Alam and Johor Bahru on the Peninsular Malaysia, and Kuching on the island of Borneo. A three-year (2007-2009) database was statistically analysed using the Pearson correlation and MLR methods. The results obtained through these analyses show that at all the three stations, PM10 has a negative relationship with relative humidity and wind speed, but a positive relationship with ambient temperature. Said Munir et al. (2013) modelled PM10 concentrations with the aid of meteorological variables and traffic-related air pollutant concentrations, such as CO, SO2, NO, NO2 and lagged PM10, employing GAMs. They identified that meteorological variables, such as temperature and wind speed largely controlled PM10 concentrations. Arwa S. Sayegh et al. (2014) evaluated several approaches including linear, non-linear and machine learning methods for the prediction of urban PM10 concentrations in the City of Makkah, Saudi Arabia. The models employed are MLR(MLRM), Quantile Regression Model (QRM), GAM, and Boosted Regression Trees1-way (BRT1) and 2-way (BRT2). Several meteorological parameters (wind speed, wind direction, temperature, and relative humidity) and chemical species measured during 2012 are used as covariates in the models. Arie Dipareza Syafei et al. (2015) investigated the prediction of each of air pollutants as dependent variable using lag-1(30 minutes before) values of air pollutants (NO2, PM10 and O3) and meteorological factors and temporal variables as independent variables by taking into account serial error correlations in the predicted concentration. Alternative variables selection based on independent component analysis (ICA) and PCA were used to obtain subsets of the predictor variables to be imputed into the linear model. The data was taken from five monitoring stations in Surabaya City, Indonesia with data period between March-April 2002. The regression with variables extracted from ICA was the worst model for all pollutants NO2, PM10 and O3 as their residual errors were highest compared with other models. The prediction of one-step ahead 30-mins interval of each pollutant NO2, PM10, and O3 was best obtained by employing original variables combination of air pollutants and meteorological factors.

As is the case with all environmental problems, the two primary causes of air pollution in Turkey are urbanization which has been rapid since the 1950s and industrialization. Before industrialization, more than 80 % of the population lived in rural areas, now most of the population live in the cities and industrial complexes. Among the developments contributing to air pollution in the cities are incorrect urbanization for the topographical and meteorological conditions, incorrect division of urban land into lots, low quality fuel and improper combustion techniques, a shortage of green areas, an increase in the number of motor vehicles and inadequate disposal of wastes (Polat, 2009). There is an extensive literature on modelling PM10 concentrations for different cities of Turkey using different models. Mucahit Egri (1997) investigated the effects of meteorological conditions on daily SO2 and PM levels which belonged to 1996-1997 winter session in Malatya, the city in Turkey. Analyses of dataset were performed by using multiple regression technique. SO2 and PM were chosen as dependent variables, while temperature, relative humidity, air pressure, wind speed and rainfall were chosen as independent variables. Out of wind speed all of explanatory variables were significantly associated with SO2 levels. While rainfall and wind speed effects were limited on PM levels, relative humidity, air pressure and temperature variables significantly changed with it. İhsan Cicek et al. (2004) showed that by using the MLR analysis that there were moderate relations between SO2, PM10, NO, NO2, CO and the climatic factors such as temperature, wind speed and humidity for Ankara, capital of Turkey. Yucel Tasdemir et al. (2005) analyzed the temporal changes in concentrations of criteria air pollutants such as CO, NOx (NO+NO2), SO2 and PM by using meteorological factors such as air temperature, relative humidity and wind speed. All of the variables measured in the period of May 2001 and April 2003 in the city of Bursa, Turkey. Correlations among pollutant concentrations and meteorological parameters showed weak relations nearly in all dataset. Lower concentrations were observed in the summer months while higher concentrations were measured in the winter months. S. Cukurluoglu and U. Guner Bacanli (2012) investigated the relationship between monthly average concentrations of SO2 and PM10 measured in the city of Denizli, situated in the southwest part of Turkey, with meteorological factors such as temperature, precipitation, humidity, and wind velocity for the period of 1994-2009. Air pollution is one of the most important environmental problems in Denizli especially during the winter periods because of the increasing of energy consumption, usage of inappropriate fuels and topography of the city. MLR analysis method was utilized to evaluate the relationships among variables. The correlation coefficients between the meteorological dataset and the SO2 and PM10 concentrations showed small relationships. Probably, other factors (i.e., traffic loadings, fugitives) may mask the relationships. Utkan Ozdemir and Simge Taner (2014) investigated prediction capacities of MLR and ANN onto coarse PM10 concentrations. Different meteorological factors, which were hourly air temperatures, wind speed, wind direction, air pressure and relative humidity, effecting on particulate pollution were chosen for operating variables in the model analyses. Two different regions (urban and industrial) were identified in the region of Kocaeli, Turkey. Regression equations explained the effects of the meteorological factors in MLR analyses. In the ANN model, backpropagation network with two hidden layers had achieved the best prediction efficiency. ANN models displayed more accurate results compared to MLR.

With increasing energy consumption, the air pollution has become an important issue. Ankara which is the second largest city and capital of Turkey is also affected by air pollution. MLR and ANN models have been commonly used in previous studies to investigate the relationship between the concentration of air pollutants and meteorological parameters. In this study, the relationship between PM10 concentrations and meteorological parameters such as press, solar radiation, cloudiness, humidity, wind speed, temperature, rainfall has been analysed statistically for Ankara. The aim of this study is to compare the performance of four linear models: MLR, RR, PCR and PLSR both in terms of modelling and predictive abilities of the PM10 concentration levels by using meteorological parameters as explanatory variables in case of dataset having a multicollinearity problem. The rest of the paper is organized as follows. In Section 2, theoretical aspects of MLR, RR, PCR and PLSR methods are presented. Section 3 presents model validation and determination of the ideal number of components retaining in PCR and PLSR models. In Section 4, the methods are compared in terms of model fit and prediction using a real air pollution dataset of Ankara city of Turkey for the period February 2007 and the results of the applications are discussed. Conclusions are reported in Section 5.

2. Multiple Linear Regression, Ridge Regression, Principal Component Regression and Partial Least Squares Regression

In this section, MLR, RR and PCR methods are briefly outlined while PLSR method is presented in more detail. The emphasis here is on the algebraic derivation of the vector (or matrix) of coefficients in the linear regression models for these four methods. Throughout this paper, matrices are denoted by bold capital letters and vectors are denoted by bold lowercase letters.

2.1. Multiple Linear Regression

Firstly, the regression model used for this method is defined by the equation (2.1).

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{2.1}$$

Traditionally, the most frequently used method for finding $\boldsymbol{\beta}$ is the OLS. If X has a full rank of p (number of independent variables) then the OLS estimate of $\boldsymbol{\beta}$ given by equation (2.2). $\hat{\boldsymbol{\beta}}_{\mathrm{OLS}}$ gives unbiased estimates for the elements of $\boldsymbol{\beta}$. The corresponding vector of fitted values obtained as in equation (2.3) (Phatak and De Jong, 1997).

$$\hat{\boldsymbol{\beta}}_{\mathrm{OLS}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \tag{2.2}$$

$$\hat{\mathbf{y}}_{\mathrm{OLS}} = \mathbf{X}\hat{\boldsymbol{\beta}}_{\mathrm{OLS}} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \mathbf{H}_{x}\mathbf{y} \tag{2.3}$$

In order to estimate $\boldsymbol{\beta}$, by using the OLS method, requires that the X variables must be linearly independent and the number of independent variables, p, must be equal or smaller than the number of observations, n $(p \le n)$ (Trygg, 2002).

If there is a multicollinearity problem, the variance of the least squares (LS) estimator may be very large and subsequent predictions rather inaccurate. However, if insisting on unbiased estimators given up, biased methods can be used to overcome the problem of inaccurate predictions. For this reason, biased methods such as RR, PCR and PLSR are used with the consequent trade-off between increased bias and decreased variance. The idea behind PCR and PLSR methods is to discard the irrelevant and unstable information and to use only the most relevant part of the x-variation for regression. Hence, the collinearity problem could be solved that more stable regression equations and predictions obtained. There are number of ways to detect or diagnose the multicollinearity. One of the most common techniques in statistics for detecting multicollinearity is the Variance Inflation Factor (VIF). The larger the VIF value, the more serious the collinearity problem. In practice, if any of the VIF values is equal or larger than 10, there is a near collinearity. In this case, the regression coefficients are not reliable (Naes et al., 2002).

2.2. Ridge Regression

The $\hat{\boldsymbol{\beta}}_{\text{OLS}}$ estimator is an unbiased and has a minimum variance. However, when multicollinearity exists, the matrix $\mathbf{X}'\mathbf{X}$ becomes ill-conditioned (singular). Since $\text{Var}\left(\hat{\boldsymbol{\beta}}_{\text{OLS}}\right) = \sigma^2 (\mathbf{X}'\mathbf{X})^{-1}$ and the diagonal elements of $(\mathbf{X}'\mathbf{X})^{-1}$ become quite large, this makes the variance of $\hat{\boldsymbol{\beta}}_{\mathbf{OLS}}$ to be large. This leads to an unstable estimate of $\boldsymbol{\beta}$ and and some of coefficients have wrong sign. In order to prevent these difficulties of OLS, Hoerl and Kennard (1970) suggested RR as an alternative procedure to the OLS method in regression analysis, especially, multicollinearity exists. The ridge technique is based on adding a biasing constant k to the diagonal elements of $\mathbf{X}'\mathbf{X}$. The RR estimator is given by equation (2.4). Here, a standardized X is used and a small positive constant k is added to the diagonal elements of $\mathbf{X}'\mathbf{X}$ (Hoerl and Kennard, 1970; Myers, 1990; Salh, 2014).

$$\hat{\boldsymbol{\beta}}_{\mathbf{RR}} = [\mathbf{X}'\mathbf{X} + k\mathbf{I}]^{-1} \mathbf{X}'\mathbf{Y}, \quad k \geq 0 \tag{2.4}$$

Here, k is ridge parameter, I is $p \times p$ identity matrix and $\mathbf{X}'\mathbf{X}$ is the correlation matrix of independent variables. Values of k lie in the range 0-1. Note that if k=0, the ridge estimator ($\hat{\boldsymbol{\beta}}_{\mathbf{RR}}$) becomes as the OLS estimator ($\hat{\boldsymbol{\beta}}_{\text{OLS}}$) (Myers, 1990; Salh, 2014).

The trick in RR is to determine the optimum value of k for developing a predictive model. There are many procedures in the literature for determining the best value. Hoerl, Kennard and Baldwin (1975) suggested to use the criterion given in equation (2.5) in order to choose the optimum k value (Rawlings, 1988).

$$k = ps^2 / \left[\hat{\boldsymbol{\beta}}_{\mathbf{OLS}}(\mathbf{0})' \hat{\boldsymbol{\beta}}_{\mathbf{OLS}}(\mathbf{0}) \right] \tag{2.5}$$

Here, p is the the the number of regression vectors except the constant term (β_0), s^2 is the estimated residual mean of squares in OLS method. The denominator of equation (2.5) shows the sum of squares of classic OLS regression parameters $\hat{\boldsymbol{\beta}}_{\mathbf{OLS}}(\mathbf{0})$, which are calculated from centered and scaled independent variables and the constant term is excluded in the calculation. The simplest way to determine the optimum value of k is to plot the values of each $\hat{\boldsymbol{\beta}}_{\text{RR}}$ versus k (in the range 0-1), which is called as ridge trace (Myers, 1990). In ridge trace graph, a trace or a curve is formed for each of the coefficients. From the ridge trace the minimum k value that makes the $\hat{\boldsymbol{\beta}}_{RR}$ stable is could be chosen and in this chosen k value the residual sum of squares could converge its minimum value (Hoerl and Kennard, 1970; Polat, 2009). However, Van Nostrand (1980) stated that since the determination of what is the stability in ridge trace is subjective and hence the selection of k is arbitrary, there is a tendency to choose a very big value of k while choosing k based on the ridge trace. Hence, the selection of k based on equation (2.5) could be better (Rawlings, 1988).

2.2. Principal Component Regression

PCR approaches the collinearity problem from the point of view of eliminating from consideration those dimensions of the X-space that they are causing the collinearity problem (Rawlings, 1988). The aim in PCR is to obtain the number of principal components (PCs) providing the maximum variation of X, which optimizes the predictive ability of the model. PCR is actually a linear regression method in which the dependent variable is regressed on the PCs. The ith PC, $\boldsymbol{\xi}_i$, is equal to $\mathbf{X}\boldsymbol{\gamma}_i$. The weight vectors, $\boldsymbol{\gamma}_i$'s $(i = 1, \cdots, p)$ satisfy equation (2.6) with equation (2.7) as shown in below.

$$\mathbf{X}'\mathbf{X}\boldsymbol{\gamma}_i = \boldsymbol{\lambda}_i\boldsymbol{\gamma}_i, \ \ i = 1, \cdots, p \tag{2.6}$$

$$\boldsymbol{\gamma}_i'\boldsymbol{\gamma}_j = \begin{cases} 1, & i = j \\ 0, & ij \end{cases} \tag{2.7}$$

In equation (2.6) $\boldsymbol{\lambda}_i$'s are the eigenvalues of the variance–covariance matrix $\mathbf{X}'\mathbf{X}$. The $\boldsymbol{\gamma}_i$'s are the unit-norm eigenvectors of $\mathbf{X}'\mathbf{X}$. If the both sides of equation (2.6) are multiplied by $\boldsymbol{\gamma}_i'$, it is easy to obtain equation (2.8), therefore, the variance of a PC is proportional to its corresponding eigenvalue. Furthermore, a given PC is orthogonal to all other PCs shown as in equation (2.9) (Phatak and De Jong, 1997).

$$\gamma_{\mathbf{i}}'\mathbf{X}'\mathbf{X}\gamma_{\mathbf{i}} = \xi_{\mathbf{i}}'\xi_{\mathbf{i}} = \lambda_{\mathbf{i}} \tag{2.8}$$

$$\boldsymbol{\gamma}_i'\mathbf{X}'\mathbf{X}\boldsymbol{\gamma}_j = \xi_{\mathbf{i}}'\xi_{\mathbf{j}} = 0, \ i \ j \tag{2.9}$$

Since neither of the PCR or PLSR methods is invariant to changes of the scale of the x-variables, PCR operates on the centered and scaled independent variables, Z. If the units of some of the variables are changed, the estimated parameters and the regression vector also will be changed. Therefore, if x variables used in the model are measured on very different scales, it may be wise to standardize the variables before modeling (Martens and Naes, 1989; Naes et al., 2002; Rawlings, 1988). The singular value decomposition (SVD) of Z has been used in the analysis of the correlational structure of the X-space. The SVD of Z matrix is given as in equation (2.10). Here U $(n \times p)$ and V $(p \times p)$ are matrices containing the left and right eigenvectors, respectively and L1/2 is the diagonal matrix of singular values. The singular values and their eigenvectors are ordered so that, $\boldsymbol{\lambda}_1 > \cdots > \boldsymbol{\lambda}_p$. The eigenvectors are pairwise orthogonal and scaled to have unit length so that $\mathbf{U'U} = \mathbf{V'V} = \mathbf{I}$.

$$\mathbf{Z} = \mathbf{UL}^{1/2}\mathbf{V}' \tag{2.10}$$

The PCs of Z are defined as the linear functions of the $\mathbf{Z}_j$ specified by the coefficients in the column vectors of V. The first eigenvector in V defines the first PC, the second eigenvector in V defines the second PC and so on. Each PC is a linear combination of all the independent variables. The matrix of PCs can be written as in equation (2.11).

$$\mathbf{W} = \mathbf{ZV} \tag{2.11}$$

The linear model in equation (2.12) can be written in terms of the PCs W as shown in equation (2.13).

$$\mathbf{y} = \mathbf{Z}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{2.12}$$

$$\mathbf{y} = \mathbf{W}\boldsymbol{\gamma} + \boldsymbol{\varepsilon} \tag{2.13}$$

$\mathbf{VV}' = \mathbf{I}$ is used to transform $\mathbf{Z}\boldsymbol{\beta}$ into $\mathbf{W}\boldsymbol{\gamma}$ as shown in equation (2.14),

$$\mathbf{Z}\boldsymbol{\beta} = \mathbf{ZVV}'\boldsymbol{\beta} = \mathbf{W}\boldsymbol{\gamma} \tag{2.14}$$

where $\boldsymbol{\gamma} = \mathbf{V}'\boldsymbol{\beta}$ is the vector of regression coefficients for the PCs and $\boldsymbol{\beta}$ is the vector of regression coefficients for the Z's. The translation of $\boldsymbol{\gamma}$ back to $\boldsymbol{\beta}$ is as in equation (2.15). Therefore, the estimate of $\boldsymbol{\beta}$ is obtained as $\hat{\boldsymbol{\beta}} = \mathbf{V}\hat{\boldsymbol{\gamma}}$. If all PCs are used, the results of $\hat{\boldsymbol{\beta}} = \mathbf{V}\hat{\boldsymbol{\gamma}}$ are equivalent to MLR (Rawlings, 1988).

$$\boldsymbol{\beta} = \mathbf{V}\boldsymbol{\gamma} \tag{2.15}$$

2. 3. Partial Least Squares Regression

PLSR is probably the least restrictive of the various multivariate extensions of the MLR model. This flexibility allows it to be used in situations where the use of traditional multivariate methods is severely limited, such as when there are fewer observations than independent variables. In these situations, the MLR approach is not feasible due to the multicollinearity between the explanatory variables (Pires et al., 2008). PLSR originated in the social sciences, especially in economy, by Herman Wold (1966). It was first presented as an algorithm akin to the power method (used for computing eigenvectors). To regress the Y variables on the X variables, PLSR attempts to find new factors that will play the same role as the X variables. These new factors often called LVs or components. Each component is a linear combination of independent variables. There are some similarities of PLSR with the PCR. In both methods, some attempts are made to find some factors that will be regressed with the Y variables. The basic difference is, while PCR uses only the variation of X to construct new factors, PLSR uses both the variation of X and Y to construct new factors which will play the role of independent variables (Phatak and De Jong, 1997). In PLSR LVs are chosen in such a way that provides maximum correlation with dependent variable. Thus, PLSR model contains the smallest necessary number of factors (Pires et al., 2008).

PLSR derives its usefulness from its ability to analyze data with many, noisy, collinear and even incomplete variables in both X and Y. PLSR method became popular first in chemometrics due in part to Herman's son Svante Wold (2001). Chemometrics (i.e., computational chemistry) is the use of statistical and mathematical procedures to extract information from chemical and physical data. Since its introduction into chemometrics as a tool for solving regression problems with highly collinear predictor variables, PLSR has become a common, unless standard regression method (Martens and Naes, 1989). The success of PLSR in chemometrics resulted in a lots of applications in other scientific areas including bioinformatics, food research, medicine, pharmacology, social sciences and physiology etc. Much work has gone into clarifying its mathematical and statistical properties. All these efforts have helped to clarify the diverse aspects of PLSR and shed light on what was it, while considering it only as an algorithm. Although PLSR is heavily improved and used by chemometricians, it used to be overlooked by statisticians and it is considered rather an algorithm than an exact statistical model for a long time. In a regression analysis, the correlation between independent variables (multicollinearity) may pose a serious difficulty in the interpretation of which independent variables are the most influential to the dependent variables. One way to remove such multicollinearity is using PCA or Partial Least Squares (PLS) analysis. Even though these methods have their own approach, the goal is similar is to build components that are statistically independent with each other. In regression analysis, this is particularly very useful and become good input as predictors in a regression model since they optimize spatial patterns and remove complexity due to multicollinearity (Syafei et al., 2015). Hence, the use of PLSR method for regression problems began in the early 80's. Since that time PLSR is also used as a multivariate regression method, which gives stable estimates especially in case of multicollinearity problem. Hence, PLSR method also used in modelling air pollution data. For Example, J. C. M. Pires et al. (2008) compared five linear models to predict the daily mean PM10 concentrations. The linear models proposed were: MLR, PCR, independent component regression (ICR), quantile regression (QR) and PLSR. The study was based on dataset from an urban site in Oporto Metropolitan Area and the analysed period was from January 2003 to December 2005. The linear models evaluated with two datasets of different sizes belonging to the analysed period. Environmental data (SO2, CO, NO, NO2 and PM10 concentrations) and meteorological data (temperature, relative humidity and wind speed) were used as PM10 predictors. During the training step, QR presented the lowest residual errors for the two datasets. ICR was the worst model using the large dataset. MLR, PCR and PLSR presented the similar results for both datasets. During the test set, ICR and QR showed bad performance, while MLR, PCR and PLSR presented similar results using the larger dataset. For the smaller dataset, the models that remove the correlation of the variables (PCR, ICR and PLSR) presented better results than MLR and QR. ICR was the linear model with the lowest value of residual error. Concluding, the dataset size is also an important parameter for the evalaution of the models concerning the prediction of variables. The prediction of the daily mean PM10 concentrations was more efficient when using ICR for the smaller dataset and PLSR for the larger datasets. It is obvious from this study that PLSR is one of the forefront methods in terms of prediction ability also for air pollution datasets studies since it removes the correlation of the variables.

2.3.1. Partial Least Squares Regression Model

The PLSR model finds a few new variables, which are estimates of the LVs or their rotations. These new variables are called X-scores and denoted by $\mathbf{t}_{\mathrm{a}}\left(\mathrm{a}=1,\cdots,\mathrm{A}\right)$. The X-scores are few (A in number) and orthogonal. They are estimated as linear combinations of the original variables $\mathbf{x}_{\mathrm{k}}$ with the weights $\mathbf{w}^{*}_{\mathrm{ka}}\left(\mathrm{a}=1,\ldots,\mathrm{A}\right)$. These weights sometimes denoted by $\mathbf{r}_{\mathrm{ka}}$. Formulas are shown both in element and matrix form (the latter in parentheses):

$$\mathbf{t}_{\mathrm{ia}}=\sum_{\mathrm{k}}\mathbf{w}^{*}_{\mathrm{ka}}\mathbf{x}_{\mathrm{ik}} \qquad \left(\mathbf{T}=\mathbf{X}\mathbf{W}^{*}\right) \tag{2.16}$$

The X-scores (ta’s) have the following properties:

a. They are multiplied by the loadings, $\mathbf{p}_{\mathrm{ak}}$ which are good summaries of X, so that the X-residuals, $\mathbf{e}_{\mathrm{ik}}$ in equation (2.17) are small.

$$\mathbf{x}_{\mathrm{ik}}=\sum_{\mathrm{a}}\mathbf{t}_{\mathrm{ia}}\mathbf{p}_{\mathrm{ak}}+\mathbf{e}_{\mathrm{ik}} \qquad \left(\mathbf{X}=\mathbf{T}\mathbf{P}'+\mathbf{E}\right) \tag{2.17}$$

With multivariate Y, the corresponding Y-scores (ua) are multiplied by the weights, cam which are good summaries of Y, so that the residuals, $\mathbf{g}_{\mathrm{im}}$ in equation (2.18) are small.

$$\mathbf{y}_{im} = \sum_a \mathbf{u}_{ia}\mathbf{c}_{am} + \mathbf{g}_{im} \qquad (\mathbf{Y} = \mathbf{UC}' + \mathbf{G}) \tag{2.18}$$

b. the X-scores are good predictors of Y, shown as in equation (2.19).

$$\mathbf{y}_{im} = \sum_a \mathbf{c}_{ma}\mathbf{t}_{ia} + \mathbf{f}_{im} \qquad (\mathbf{Y} = \mathbf{TC}' + \mathbf{F}) \tag{2.19}$$

The Y-residuals, $\mathbf{f}_{im}$, express the deviations between the observed and modeled dependent variables and form the elements of the Y-residual matrix, F. Using the equation (2.16), equation (2.19) can be rewritten to look as a multiple regression model as in equation (2.20). So that the PLSR coefficients, $\mathbf{b}_{mk}(\mathbf{B})$, could be written as in equation (2.21).

$$\mathbf{y}_{im} = \sum_a \mathbf{c}_{ma}\sum_k \mathbf{w}^*_{ka}\mathbf{x}_{ik} + \mathbf{f}_{im} = \sum_k \mathbf{b}_{mk}\mathbf{x}_{ik} + \mathbf{f}_{im} \qquad (\mathbf{Y} = \mathbf{XW}^*\mathbf{C}' + \mathbf{F} = \mathbf{XB} + \mathbf{F}) \tag{2.20}$$

$$\mathbf{b}_{mk} = \sum_a \mathbf{c}_{ma}\mathbf{w}^*_{ka} \qquad (\mathbf{B} = \mathbf{W}^*\mathbf{C}') \tag{2.21}$$

After each component, a, the X-matrix is deflated by subtracting $\mathbf{t}_a\mathbf{p}'_a$ from X. This makes the PLSR model alternatively be expressed by weights wa referring to the residuals after previous dimension, $\mathbf{E}_{a-1}$, instead of relating to the X-variables themselves. Therefore, instead of equation (2.16), equation (2.22) can be written. The relation between the two weights is given as in equation (2.25) The Y-matrix can also be deflated by subtracting $\mathbf{t}_a\mathbf{c}'_a$, but this is not necessary. The results are equivalent with or without Y-deflation (Wold et al., 2001).

$$\mathbf{t}_{ia} = \sum_k \mathbf{w}_{ka}\mathbf{e}_{ik,a-1} \qquad (\mathbf{t}_a = \mathbf{E}_{a-1}\mathbf{w}_a) \tag{2.22}$$

$$\mathbf{e}_{ik,a-1} = \mathbf{e}_{ik,a-2} - \mathbf{t}_{i,a-1}\mathbf{p}_{a-1,k} \qquad (\mathbf{E}_{a-1} = \mathbf{E}_{a-2} - \mathbf{t}_{a-1}\mathbf{p}'_{a-1}) \tag{2.23}$$

$$\mathbf{e}_{ik,0} = \mathbf{x}_{ik} \qquad (\mathbf{E}_0 = \mathbf{X}) \tag{2.24}$$

$$\mathbf{W}^* = \mathbf{W}(\mathbf{P}'\mathbf{W})^{-1} \tag{2.25}$$

2.3.2. Partial Least Squares Regression Algorithms

There are several ways to calculate PLSR model parameters. Perhaps the most intuitive method known as Non-Linear Iterative Partial Least Squares (NIPALS), which is also called as a classical algorithm. NIPALS calculates scores, T and loadings, P and an additional set of vectors known as weights, W (with the same dimensionality as the loadings P). The addition of weights in PLSR is required to maintain orthogonal scores (Wold et al., 2001). The simple NIPALS algorithm of Wold et al. (1984) is shown as below. It starts with optionally transformed, scaled, centered data (X and Y) and proceeds as follows. If there is a single y-variable, the algorithm is non-iterative.

A. Get a starting vector of u, usually one of the Y columns. With a single y, u=y.

B. The **X**-weights, **w**: $\mathbf{w} = \mathbf{X}'\mathbf{u} / \mathbf{u}'\mathbf{u}$. Scale **w** to be of length one.
C. Calculate **X**-scores, **t**: $\mathbf{t} = \mathbf{X}\mathbf{w}$.

D. The **Y**-weights, **c**: $\mathbf{c} = \mathbf{Y}'\mathbf{t} / \mathbf{t}'\mathbf{t}$.

E. Finally, an updated set of **Y**-scores, **u**: $\mathbf{u} = \mathbf{Y}\mathbf{c} / \mathbf{c}'\mathbf{c}$.

F. Convergence is tested on the change in **t**, i.e., $\left\|\mathbf{t}_{\text{old}} - \mathbf{t}_{\text{new}}\right\| / \left\|\mathbf{t}_{\text{new}}\right\| < \varepsilon$. Where ε is very small positive number, e.g., 10^{-6} or 10^{-8}. If convergence has not been reached, return to B, otherwise continue with G and then A. If there is only one **y**-variable, the procedure converges in a single iteration and one proceeds directly with G.

G. Remove (deflate) the present component from **X** and **Y** and then use these deflated matrices as new **X** and **Y**, while computing the next component. Here the deflation of **Y** is optional, the results are equivalent whether **Y** is deflated or not.

X-loadings: $\mathbf{p} = \mathbf{X}'\mathbf{t} / (\mathbf{t}'\mathbf{t})$
Y-loadings: $\mathbf{q} = \mathbf{Y}'\mathbf{u} / (\mathbf{u}'\mathbf{u})$
Regression (**u** upon **t**): $\mathbf{b} = \mathbf{u}'\mathbf{t} / (\mathbf{t}'\mathbf{t})$
Residual matrices: $\mathbf{X} \rightarrow \mathbf{X} - \mathbf{t}\mathbf{p}'$ and $\mathbf{Y} \rightarrow \mathbf{Y} - \mathbf{b}\mathbf{t}\mathbf{c}'$

H. Continue with next component (back to step A) until CV (see below) indicates that there is no more significant information in **X** about **Y**.

The next set of iterations of algorithm starts with the new X and Y matrices as the residual matrices from the previous iteration. The iterations can continue until a stopping criteria is used or X becomes the zero matrix (Höskuldsson, 1988; Wold et al., 2001).

The major drawback of NIPALS algorithm is that the columns of score matrix T are obtained as linear combinations of deflated data matrix X. Since one looses sight of what is in the deflated data, the interpretation of the components becomes complicated. Hence, the other significant algorithm for PLSR, Straightforward Implementation of a Statistically Inspired Modification of the Partial Least Squares Method (SIMPLS), was proposed by Sijmen De Jong (1993). This algorithm aims to derive the PLS factors, T, directly as linear combinations of the original X variables that causing many advantages for it. Firstly, the deflation of the data matrices X and/or Y as in NIPALS becomes unnecessary, which may result in faster computation and less memory requirements. Secondly, all factors are equally easy to interpret because of occurring as simpler linear combinations of the original variables. Another advantage is that the final PLSR model can be derived very easily when the factors are already expressed as linear combinations of the original variables. SIMPLS deflates the cross-covariance matrix, $\mathbf{S}_{xy} \propto \mathbf{X}'\mathbf{Y}$, whereas NIPALS deflates the original data matrix X to obtain orthogonal components. Hence, this method is very non-intuitive but very fast and accurate. It gives the exact same result as NIPALS for univariate y, but a slightly different solution for multivariate Y. The SIMPLS algorithm is a very fast PLSR algorithm for all kinds of shapes of data matrices. This approach does not always give the same model as NIPALS, but the difference is very small and for most cases not significant (De Jong, 1993; Lindgren and Rännar, 1998).

3. Model Validation and Determination of the Ideal Number of Components Retaining in PCR and PLSR Models

In literature there are several measures of a model's fit to the data and predictive power. In all of the measures investigated, to estimate the average deviation of the model from the data is attempted. The root mean square error (RMSE) is a measure of how well model fits the data. It is defined as,

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2}{n}} \tag{2.26}$$

where $\hat{y}_i$ are the values of the predicted variable when all samples included in the model construction and n is the number of observations (Wold et al., 2001).

Root-mean-square error of cross-validation (RMSECV), however, is a model's ability to predict new samples that is in contrast to RMSE. RMSECV is defined as RMSE, the only difference is that $\hat{y}_{CV,i}$ are predictions for samples not included in the model formulation. In RMSECV, $\hat{y}_{CV,i}$ contains the values of the y variable that are estimated by cross-validation (CV) (Naes et al., 2002; Wise et al., 2006).

$$\mathrm{RMSECV} = \sqrt{\frac{\mathrm{PRESS}}{n}} = \sqrt{\frac{\sum_{i=1}^{n}\left(y_i - \hat{y}_{CV,i}\right)^2}{n}} \tag{2.27}$$

CV is a method used for selecting the optimal number of components, which maximize model's predictive ability, for PCR and PLSR methods. Hence, as shown in equation (2.27) RMSECV is related to the prediction sum of squares (PRESS) value for the number of components (k) included in the model for both of PCR and PLSR methods. Of course, the exact value of RMSECV depends not only on k for both PCR and PLSR methods, but for all the methods (MLR, RR, PCR, PLSR) on how the CV subsets (test sets) were formed. PRESS statistics is a measure, which assesses model's validation and predictive ability. In general, the smaller the PRESS value, the better the model's predictive ability (Naes et al., 2002; Wise et al., 2006; Polat, 2009).

The performance of CV approach naturally could be changed in connection with the number of removed observations. For a given dataset, CV involves a series of experiments, hereby called ‘sub-validation experiments’, each of which involves the removal of a subset of objects from a dataset (the test/validation set), construction of a model using the remaining objects in the dataset (the model building set) and subsequent application of the resulting model to the removed objects. This way, each sub-validation experiment involves testing a model with objects that they are not used to build the model. A typical CV procedure usually involves more than one sub-validation experiment, each of which involves the selection of different subsets of samples for model building and model testing. As there are several different modeling methods, there are also several different CV methods and these vary with respect to how the different sample subsets are selected for these sub-validation experiments. For example, Leave one out cross-validation (LOOCV) performs CV each time leaving out one observation out of the model formulation and predicting once. LOOCV method is generally reserved for small datasets (n not greater than 20). When using LOOCV with more than 20 samples, the CV may be biased and report errors lower than might be expected. Hence, it is preferred to consider using another CV method in these cases. In this study, the number of components retaining in the PCR and PLSR models is determined by RMSECV values and ‘venetian blinds CV method’ is used for calculating these values. The ‘venetian blinds CV method’ is an alternative CV method given in MATLAB PLS_Toolbox. In this method, by using ‘venetian blinds’ approach the dataset is divided into several subgroups. In this approach, the data is not divided to random parts. n is the total number of objects in the dataset and s, which must be less than n/2, is the number of data splits specified for the CV procedure. For ‘venetian blinds CV method’, each subset is determined by selecting every sth object in the dataset, starting at objects numbered 1 through s. Hence, s shows the number of subsets. Then, one of the obtained s subsets is left out for validaton analysis and remained (s-1) subsets are used for constructing the model. The prediction performance of obtained model is evaluated by using this removed sth subset. This process is repeated until all the subsets is left out once and in order to obtain the last validality measure the mean of the obtained prediction errors are calculated. This method is simple and easy to implement, and generally safe to use if there are relatively many objects that are already in random order. The number of subsets for this method is automically selected in PLS_Toolbox by taking the nearest integer to the square root of the number of observation in the dataset (Wise et al., 2006, Eigenvector website). For Example, since in this study we use the period of February 2007 there are 28 observations in the analysis. When we choose 'venetian blinds CV method' in PLS_Toolbox this 28 observations is partitioned in to 5 subsets as shown in below:

Subset 1: 1, 6, 11, 16, 21, 26
Subset 2: 2, 7, 12, 17, 22, 27
Subset 3: 3, 8, 13, 18, 23, 28
Subset 4: 4, 9, 14, 19, 24
Subset 5: 5, 10, 15, 20, 25

Since it directly addresses the collinearity problem, PCR is less susceptible to overfitting than MLR. However, one could still overfit a PCR model, through the retention of too many PCs. Therefore, an important part of PCR is the determination of the optimal number of PCs to retain in the model. For problems in which there are more observations than variables a PCR model, which has the maximum possible number of PCs (which is equal to the number of variables), becomes identical to the MLR model built using the same set of dependent variables. In a sense, the PCR model converges to the MLR model as PCs are added. However, in real applications, it is almost always the case that the optimal number of PCs retaining in the model is much fewer than the number of original independent variables (Rawlings, 1988). Hence, the optimal number of components retaining in PCR/PLSR model depends on the specific objectives of the modeling project, but it is typically the number at which the addition of another component does not greatly improve the prediction performance of the model (Wise et al., 2006). In practice, PCR and PLSR generally have similar performance; the main advantage of PLSR is often computational speed. In addition, PLSR models generally require fewer factors than PCR models for the same set of data, hence, PLSR has advantages in both model implementation and model interpretation. Recalling that each PC is constrained to be orthogonal to all previous ones. This constraint can make interpretation of the loadings and scores for PCR progressively more difficult as the number of components increases (Wise et al., 2006).

In this study, the Wilcoxon signed rank test is also employed to compare the predictive performance of pairs of four models (MLR, RR, PCR, PLSR) calculated with the air pollution dataset. Let the set of prediction errors associated with the first model is denoted as $\mathbf{e_1} = \{e_{11}, e_{12}, \cdots, e_{1n}\}$. Let the set of prediction errors associated with the second model is denoted as $\mathbf{e_2} = \{e_{21}, e_{22}, \cdots, e_{2n}\}$. The prediction errors are assumed to be independent within a set. Also, let $\mathbf{a_j} = \{a_{j11}, a_{j2}, \cdots, a_{jn}\} = \{|e_{j1}|, |e_{j2}|, \cdots, |e_{jn}|\}$ for j=1,2 represent the absolute values of the prediction errors. The set of n paired comparisons between matched errors (i.e. e_{1i}

and e_{2i}) is used as the basis for making statistical inference about the relative performance of two competing models. Hypothesis testing using non-parametric statistical methods (sign test and Wilcoxon signed rank test) is proposed as a way to statistically compare the two sets of prediction errors and hence determine whether one model significantly outperformed the other for the dataset of interest. For each observation, preference is given to the model that produces the smallest absolute value of the prediction error. A very simple method for making pairwise comparisons is the sign test. However, more powerful (and complex) method for comparison is the Wilcoxon signed rank test (Lehmann, 1975; Thomas, 2003). Because while the sign test considers only the sign of the differences between absolute errors (i.e. $\operatorname{sign}(a_{1i} - a_{2i})$), the Wilcoxon signed rank test considers the magnitude of those differences. Let, $f_i = |a_{1i} - a_{2i}|$ and r_i be the integer rank (from 1 to n) of f_i among $\{f_1, f_2, \cdots, f_n\}$. Thus $\{r_1, r_2, \cdots, r_n\}$ is a permutation of the integers from 1 to n. Also, let $g_i = \operatorname{sign}(a_{1i} - a_{2i})$. Let $d = \sum r_i g_i = U - V$, where U is the sum of the positive ranks (i.e. when $g_i = 1$) and V is the sum of the negative ranks (i.e. when $g_i = -1$). Preference for model 2 is exhibited when d>0. On the other hand, preference for model 1 is exhibited when d<0. Noting that the equation (2.28), the Wilcoxon signed rank statistic is given as in equation (2.29). Under the assumption that the two models are of equal quality, the expected value and variance of V_n are showed as in equation (2.30) and equation (2.31), respectively (Thomas, 2003).

$$U + V = n(n+1)/2 \tag{2.28}$$

$$V_n = \frac{n(n+1)/2 - d}{2} \tag{2.29}$$

$$E(V_n) = \frac{n(n+1)}{4} \tag{2.30}$$

$$V(V_n) = \frac{n(n+1)(2n+1)}{24} \tag{2.31}$$

For the general case with an arbitrarily large value of n the exact distribution of Vn is generally unavailable owing to the prohibitive enumeration required. For purposes of statistical inference,

$$z = \frac{V_n - E(V_n)}{\sqrt{Var(V_n)}} = \frac{V_n - \frac{n(n+1)}{4}}{\sqrt{\frac{n(n+1)(2n+1)}{24}}} \tag{2.32}$$

equation (2.32) is approximately normally distributed with mean zero and variance one (Lehmann, 1975; Thomas, 2003). Thus, assuming that the two models are of equal quality, the probability of observing a smaller value for z is $\Phi(z)$, where $\Phi(.)$ is the cumulative normal distribution. If $\Phi(z)$ is sufficiently small (say less than 0.05), then we conclude that the first model has outperformed the second model. If $\Phi(z)$ is sufficiently large (say greater than 0.95), then we conclude that the second model has outperformed the first model (Thomas, 2003).

4. Applications on a Real Airpollution Dataset

Considering the studies and thesis about air pollution, it is noticed that the meteorological parameters affecting the air pollution. Therefore, in this study, a real dataset of meteorological parameters measured by The Rebuplic of Turkey Ministry of Environment and Forestry Turkish State Meteorological Service used. In addition, PM10 measured by The Rebuplic of Turkey Ministry of Environment and Forestry Refik Saydam Hygiene Center used also. Ankara city is taken as the study area for the dataset. In the dataset, February 2007 considered as the period. The dataset consists of eleven meteorological parameters used as independent variables and PM10 used as the dependent variable. The air pollution dataset for this study is given in Table 1. For the dataset, eleven of the meteorological parameters that affect the forming of the PM10 (µg/m3) are given as below immediately under the Table 1 (Polat, 2009).

Table 1. Air pollution dataset for the period February 2007.

Obs. No	PM_{10} (µg/m^3)	PRESS (mb)	SOLAR (cal/cm^2)	HUMIDITY (%)	WIND (m/sn)	MIN (°C)	MAX (°C)	AVG (°C)	RAIN (mm)	FIRST-CLOUD (8/8)	SECOND-CLOUD (8/8)	THIRD-CLOUD (8/8)
1	29.00	915.5	107.25	65.3	3.67	-4.9	5.0	0.2	0.0	6.00	1.00	5.4
2	26.00	911.0	333.95	86.1	7.78	0.4	2.4	1.1	3.8	8.00	8.00	7.0
3	21.50	908.4	396.59	80.4	7.67	-2.6	0.0	-1.3	7.2	8.00	7.00	5.4
4	19.00	905.9	322.89	80.3	3.00	-3.0	0.0	-2.1	0.4	8.00	7.00	8.0
5	48.00	907.7	167.33	73.2	6.33	-8.5	0.6	-3.8	0.2	3.00	5.00	5.4
6	19.40	908.1	301.69	62.4	6.00	-6.8	2.0	-2.4	0.0	4.86	5.00	5.4
7	52.00	917.4	108.72	69.5	6.00	-9.8	2.6	-3.6	0.0	4.86	5.00	5.4
8	58.25	918.5	132.32	70.7	3.00	-6.6	5.9	-1.3	0.0	2.00	4.83	5.4
9	54.32	916.7	256.67	69.1	5.00	-4.7	7.1	0.8	0.0	6.00	5.00	5.4
10	76.00	915.8	120.23	66.2	4.00	-3.7	9.6	2.4	0.0	2.00	4.83	5.4
11	77.00	915.7	140.12	71.7	4.33	-2.6	8.4	1.7	0.0	4.00	6.00	4.0
12	93.50	912.8	241.17	73.5	4.25	-2.5	7.6	2.0	0.0	4.00	6.00	2.0
13	79.50	910.1	329.90	75.3	3.00	-0.1	9.0	3.6	0.0	7.00	7.00	4.0
14	68.25	907.4	298.43	85.5	5.00	1.5	8.6	3.7	0.0	6.00	7.00	3.0
15	54.32	909.7	233.63	80.1	3.25	1.9	9.7	5.8	4.8	4.86	6.00	6.0

16	81.00	913.9	447.32	78.9	3.50	1.3	9.1	5.2	0.0	2.00	6.00	5.4
17	49.00	914.3	138.49	71.9	7.00	0.6	11.3	5.8	0.0	4.86	3.00	5.4
18	48.50	917.7	483.36	67.1	5.33	2.3	8.8	4.8	0.0	6.00	6.00	6.0
19	39.50	915.8	167.67	63.4	6.00	-0.4	11.1	5.7	0.0	4.86	2.00	5.4
20	54.32	914.6	117.98	66.6	3.50	1.4	13.8	7.1	0.0	3.00	1.00	5.4
21	88.50	913.8	126.86	65.8	3.00	1.1	14.7	7.9	0.0	4.86	2.00	5.4
22	82.00	913.9	168.46	58.6	4.00	3.1	12.5	7.8	0.0	6.00	2.00	5.4
23	50.50	908.6	302.18	66.1	3.67	1.6	12.7	6.6	0.0	7.00	7.00	7.0
24	24.50	912.3	99.99	54.0	9.00	1.5	9.0	3.6	0.0	2.00	2.00	5.4
25	32.00	913.1	91.88	44.4	4.79	-6.4	6.8	-0.7	0.0	4.86	4.83	5.4
26	55.50	911.6	100.11	39.0	4.50	-5.9	8.7	1.4	0.0	4.00	3.00	5.4
27	83.00	910.0	110.80	41.8	5.00	-3.7	11.9	4.5	0.0	2.00	4.83	5.4
28	56.50	908.5	640.32	78.0	2.50	3.0	8.0	5.3	0.0	6.00	7.00	7.0

PRESS: Daily average press (mb)
SOLAR: Daily average solar radiation (cal/cm2)
HUMIDITY: Daily average humidity (%)
WIND: Daily average wind speed (knot) (1knot=1.852 km/hour=0.514 m/seconds)
MIN: Daily minimum temperature (℃)
MAX: Daily maximum temperature (℃)
AVG: Daily average temperature (℃)
RAIN: Daily total rainfall (mm)
FIRSTCLOUD: Cloudiness at 6.00 am (8/8)
SECONDCLOUD: Cloudiness at 12.00 am (8/8)
THIRDCLOUD: Cloudiness at 18.00 am (8/8)

Firstly, MLR analysis applied on the dataset and it is found that the model obtained by using OLS method is significant with a probability of 95% (F=8.83; p=0.000). According the MLR analysis, 85.9 % of variation occurring in PM10 variable is explained by these eleven meteorological parameters.

Even though the MLR model fits the data well, multicollinearity may severely prohibit quality of the prediction. Table 2 shows that all independent variables except WIND, MIN, AVG, TWOCLOUDINESS and THREECLOUDINESS are not significant as an indicator of multicollinearity problem. The existence of collinearity problem also could be seen by examining the VIF values. The VIF values for MIN, MAX and AVG are 17.8, 71 and 102.9, respectively. Hence, there is a near-collinearity problem for this dataset.

Table 2. The estimated regression coefficients and VIF values for the MLR model.

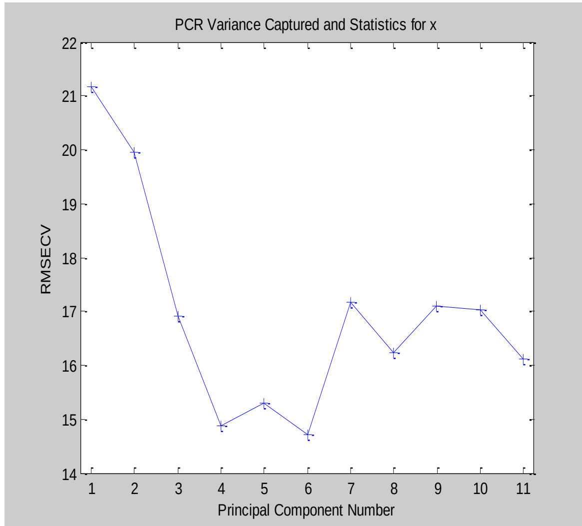
Model	Coefficients	Standart Error of Coefficients	T	P	VIF
Constant	-975.4	722.1	-1.35	0.196	

PRESS	1.2163	0.8122	1.50	0.154	1.9
SOLAR	-0.06034	0.03708	-1.63	0.123	6.0
HUMIDITY	-0.0805	0.3221	-0.25	0.806	3.4
WIND	-4.516	1.701	-2.66	0.017	1.8
MIN	-6.285	2.378	-2.64	0.018	**17.8**
MAX	-8.641	4.300	-2.01	0.062	**71.0**
AVG	18.501	6.122	3.02	0.008	**102.9**
RAINFALL	-2.910	1.870	-1.56	0.139	2.3
FIRSTCLOUDINESS	-2.834	1.462	-1.94	0.070	1.7
TWOCLOUDINESS	7.454	2.202	3.38	0.004	4.6
THREECLOUDINESS	-9.610	2.218	-4.33	0.001	1.5

Although MLR model fits to dataset well, since it has a near-collinearity problem the predictive ability of it could be not well. In that case, by using biased methods such as RR, PCR and PLSR, better models in terms of predictive ability could be obtained that cope with multicollinearity. On this dataset, by using MATLAB PLS_Toolbox MLR, RR, PCR and PLSR methods are applied. Then the models obtained by these four methods are compared in terms of their fitting ability to this dataset by using their calculated RMSE values and in terms of predictive ability for future datasets by using their RMSECV values and additionally Wilcoxon signed rank test results.

The number of PCs/LVs giving the minimum RMSECV value is chosen as the optimal for the models. Since the model having less components and small RMSECV value always preferred, as seen from Figure 1 the PCR model with four PCs have been chosen. Table 3 presents the percent X and Y variance captured by the PCR model. For the optimal number of PCs in PCR 77.08 % of the variance is captured by the new predictors. These four PCs could explain 69.37 % of the variation in the dependent variable.

Figure 1. RMSECV values for PCR and PLSR models on the air pollution dataset.



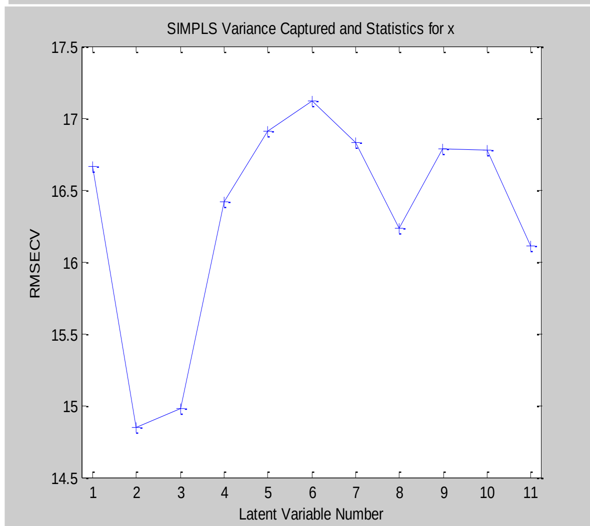


Table 3. The percent variance captured by PCR model for the air pollution dataset

	X-block		Y-block	
PC	PC	Cum	PC	Cum
1	32.44	32.44	19.14	19.14
2	24.43	56.87	14.00	33.13
3	10.69	67.56	25.44	58.57
4	**9.51**	**77.08**	**10.81**	**69.37**
5	6.96	84.04	1.55	70.93
6	5.18	89.22	0.60	71.53
7	4.73	93.94	2.35	73.88

8	3.40	97.35	3.04	76.92
9	2.11	99.46	1.09	78.01
10	0.49	99.95	1.15	79.15
11	0.05	100.00	6.71	85.86

By default, MATLAB PLS_Toolbox uses SIMPLS algorithm for PLSR because of it is speed. However, the NIPALS algorithm could be used as an option by using the menus. Both of the algorithms give same results in terms of RMSE and RMSECV statistics and the same figure for determination of the numbers of components retaining in the PLSR model. Therefore, only the results of SIMPLS algorithm are given in this study. The number of LVs retaining in the PLSR model is determined as a same manner in PCR model. As seen from Figure 1, the maximum number of LVs retaining in the PLSR model could be chosen as two. Table 4 presents the percent X and Y variance captured by the PLSR model. For the optimal number of LVs in PLSR 44.14 % of the variance is captured by the new predictors. These two LVs could explain 74.48 % of the variation in the dependent variable.

Table 4. The percent variance captured by PLSR model for the air pollution dataset

	X-block		Y-block	
LV	LV	Cum	LV	Cum
1	27.60	27.60	58.16	58.16
2	**16.54**	**44.14**	**16.32**	**74.48**
3	21.67	65.80	2.28	76.76
4	5.99	71.79	1.73	78.49
5	6.91	78.70	0.52	79.02
6	5.14	83.85	0.70	79.72
7	4.20	88.05	0.94	80.66
8	3.80	91.84	0.83	81.49
9	2.96	94.80	0.68	82.17
10	2.49	97.30	2.09	84.25
11	2.70	100.00	1.61	85.86

While comparing the models, fit and prediction are different sights of a model performance. For example, whether the aim is the selection of a model that best fits the dataset, then the model having the smallest RMSE value should be chosen. If the aim is the selection of a model giving the best prediction, instead of RMSE a different measure must be used. The best way of controlling a model's predictive ability is of course validation of that model on a new validation set. However, an independent and representative validation set is rare. In the absence of a real validation set, one reasonable way of model validation is given by CV, which simulates how well the model predicts new data. Hence, in this study, RMSECV values have been used in order to compare the methods in terms of prediction performance.

As seen from Table 5, the smallest RMSE values belong to MLR and RR models, respectively. Under the condition of no collinearity in the independent variables, MLR and RR models fit the data better than PCR and PLSR models. Due to the existence of the collinearity in the dataset used, this interpretation would not be true at all. For comparison of models intended for prediction, it is inadequate to look just at model fit. Hence, RMSECV values for all of the methods must be considered also. The best models in terms of RMSECV values are PLSR and PCR models, respectively. Therefore, MLR and RR models could be considered as the best models for fitting to dataset, however, PLSR and PCR models could be considered as the best models for prediction.

Table 5. RMSE and RMSECV values of four models for the air pollution dataset

	MLR	RR	PCR (4 PCs)	PLSR (2 LVs)
RMSE	8.27317	8.9053	12.1759	11.1149
RMSECV	16.1115	15.1505	14.8883	14.8503

Moreover, Wilcoxon signed rank test could be used to compare the predictive performance of pairs of these four models. Therefore, in terms of prediction ability the difference between pairs of models (especially MLR-RR and PCR-PLSR pairs) could be more clearly distinguishable. The results of the full pairwise model comparison made by Wilcoxon test show that in terms of predictive ability PCR and PLSR models are equivalent ($\Phi(z) = 0.2921$ for PLSR compared to PCR). PLSR model, compared with the MLR and RR models ($\Phi(z) = 0.0055$ for PLSR compared to MLR, $\Phi(z) = 0.0052$ for PLSR compared to RR), is one of two models having the best predictive ability. PCR model, compared with the MLR and RR models ($\Phi(z) = 0.0059$ for PCR compared to MLR, $\Phi(z) = 0.0049$ for PCR compared to RR), is the other one of two models having the best predictive ability. However, MLR and RR models are equivalent ($\Phi(z) = 0.2617$ for MLR compared to RR). These results are also supported the resuts of Table 5 that PLSR and PCR models could be considered as the best models for prediction.

The regression coefficients obtained for the four methods are given as shown in Table 6. From Table 6, it is obvious that the signs of the eleven coefficients are same for both of PCR and PLSR methods and these two methods show close performance.

Table 6. The estimated coefficients for MLR, RR, PCR (4PCs), PLSR (2LVs) models

Independent Variables	$\hat{\beta}_{MLR}$	$\hat{\beta}_{RR}$	$\hat{\beta}_{PCR}$	$\hat{\beta}_{PLSR}$
PRESS	0.1933	0.11461	0.1399	0.0604
SOLAR	-0.3746	-0.15717	0.0625	0.0340
HUMIDITY	-0.0432	0.04717	0.1661	0.1040
WIND	-0.3362	-0.30595	-0.2705	-0.3363
MIN	-1.0493	-0.59285	0.0741	0.0521
MAX	-1.5921	-0.40279	0.1746	0.2439
AVG	2.8817	1.3648	0.1165	0.1836
RAINFALL	-0.2234	-0.10389	-0.2146	-0.1129
FIRSTCLOUDINESS	-0.2392	-0.20957	-0.1657	-0.1807
TWOCLOUDINESS	0.6816	0.47759	0.0962	0.2074
THREECLOUDINESS	-0.4980	-0.43416	-0.4107	-0.3990

The actual and predicted values of dependent variable PM10 for MLR, RR, PCR and PLSR models are given in Table 7. It is clear from Table 7 that MLR and RR methods show close performances while PCR and PLSR methods give close results in terms of prediction values. As we mentioned before, MLR and RR methods are the best ones for fitting to this dataset. Hence, these two methods (especially MLR method) predict y values better than the other ones for this dataset. However, PCR and PLSR coefficients given in Table 6 must be chosen for prediction of PM10 values in future datasets. Because both of these two methods give lower RMSECV values and additionally, pairwise comparisons of these two ones with MLR and RR methods by using Wilcoxon signed rank test showed their superiority in terms of predicton ability again.

Table 7. Actual and predicted values of dependent variable (PM_{10}) for MLR, RR, PCR (4PCs), PLSR (2LVs) models.

y	$\hat{y}_{MLR}$	$\hat{y}_{RR}$	$\hat{y}_{PCR}$	$\hat{y}_{PLSR}$
29.00	39.6920	37.9832	48.6105	41.6669
26.00	26.1869	21.7363	19.1402	19.6312
21.50	13.4200	16.6788	12.8393	16.4868
19.00	18.4380	17.7519	24.4609	25.1197
48.00	38.3082	36.0248	35.9082	32.0929
19.40	31.4783	32.1400	35.5039	32.7205
52.00	45.3310	43.1841	42.2872	35.0195
58.25	59.4614	63.6016	66.4663	63.1653
54.32	47.3381	50.7827	54.2510	50.2945
76.00	71.0100	71.4882	66.4999	66.8468
77.00	74.7659	75.4985	74.3442	73.5754
93.50	96.4118	92.1142	89.0168	88.4626
79.50	75.4271	76.6705	75.5361	78.8494
68.25	71.8822	75.1283	79.0276	80.5255
54.32	66.7426	65.6979	51.8713	61.9706
81.00	83.6177	78.9910	78.9428	79.4889
49.00	53.5182	52.2636	54.7529	51.2554
48.50	50.5548	50.9031	59.6024	56.8340
39.50	57.4918	53.8575	55.6081	51.8896
54.32	59.1393	64.1725	71.1283	70.5739
88.50	71.0454	72.0675	70.8017	72.7478
82.00	66.0786	60.3443	61.9360	62.2532
50.50	57.0093	59.7547	51.3824	59.1808
24.50	19.9835	26.4360	40.9282	36.2788
32.00	43.3289	46.5406	42.4310	43.5424
55.50	50.8415	51.0156	43.9300	46.1156
83.00	80.9522	74.6606	53.8564	60.8610
56.50	51.4053	53.3718	59.7965	63.4107

5. Conclusion

In this comparative study, MLR, RR, PCR and PLSR methods applied on a real air pollution dataset with multicollinearity and they have been compared from the point of view of model fit and prediction. The results show that when the model fit is considerable, MLR and RR models fit to this dataset best. The dataset used in this paper has also showed that the regression models constructed by PLSR and PCR methods have the highest predictive ability, moreover, PLSR shows equal performance with PCR even with the smaller number of components. Although PCR and PLSR methods give more analogously results in terms of RMSE and RMSECV values, it is significant to mention that the results of these methods in terms of prediction and fitting are very sensitive choosing optimal number of components properly.

References

[1] Afzali, A., Rashid, M., Sabariah, B., Ramli, M. (2014). PM_{10} Pollution: Its Prediction and Meteorological Influence in PasirGudang, Johor. *8th International Symposium of the Digital Earth (ISDE8) IOP Conf. Series: Earth and Environmental Science* **18**.

[2] Barmpadimos, I., Hueglin, C., Keller, J., Henne, S. and PrévôtA. S. H. (2011). Influence of meteorology on PM_{10} trends and variability in Switzerland from 1991 to 2008. *Atmos. Chem. Phys.* **11**, 1813–1835.

[3] Chaloulakou, A., Grivas, G., Spyrellis, N. (2003). Neural Network and Multiple Regression Models for PM_{10} Prediction in Athens: A Comparative Assessment. *J. Air Waste Manage. Assoc.* **53**, 1183–1190.

[4] Cicek, İ., Turkoglu, N., Gurgen, G. (2004). Ankara'da Hava Kirliliginin İstatistiksel Analizi. *Fırat University Journal of Social Science* **14:2**, 1-18.

[5] Cukurluoglu, S. and Bacanli, U. G. (2012). Determination of Relationship between Air Pollutant Concentrations and Meteorological Parameters in Denizli, Turkey. BALWOIS 2012 - Ohrid, Republic of Macedonia - 28 May, 2 June.

[6] De Jong, S. (1993). SIMPLS: An alternative approach to partial least squares regression. *Chemometrics and Intelligent Laboratory Systems* **18**, 251-263.

[7] Demuzere, M., Trigo, R. M., Vila-Guerau de Arellano, j. and van Lipzig, N. P. M. (2009). The impact of weather and atmospheric circulation on O_3 and PM_{10} levels at a rural mid-latitude site. *Atmos. Chem. Phys.* **9**, 2695–2714.

[8] Dominick D., Latif M. T., Juahir H., Aris, A. Z. and Zain, S. M. (2012). An assessment of influence of meteorological factors on PM and NO at selected stations in Malaysia. *Sustain. Environ. Res.* **22(5)**, 305-315.

[9] Egri, M. (1997). 1996-1997 Kıs Döneminde Malatya İl Merkezi Hava Kirliligi Parametrelerine Meteorolojik Kosulların Etkisi. Journal of Turgut Ozal Medical Center **4(3),** 265-269.

[10] Helland, I.S. (1988). On the structure of partial least squares regression. *Communications in Statistics- Simulation and Computation* **17**, 581-607.

[11] Hoerl, A. E., Kennard, R. W. (1970). Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics*, **12(1)**, 55-67.

[12] Hörmann, S., Pfeiler, B., and Stadlober E. (2005). Analysis and Prediction of Particulate Matter PM_{10} for theWinter Season in Graz. *Austrian Journal of Statistics* **34:4**, 307–326.

[13] Höskuldsson, A. (1988). PLS Regression Methods. *Journal of Chemometrics* **2**, 211-228.

[14] Lindgren, F. and Rännar, S. (1998). Alternative Partial Least-Squares (PLS) Algorithms. *Perspectives in Drug Discovery and Design* **12/13/14**, 105–113.

[15] Lehmann EL. (1975). *Nonparametrics: Statistical Methods Based on Ranks*, Holden-Day: San Francisco, CA, 120–132.

[16] Martens, H. and Naes, T. (1989). *Multivariate Calibration,* New York, Brisbane, Toronto, Singapore: John Wiley & Sons.

[17] McKendry I. G. (2002). Evaluation of Artificial Neural Networks for Fine Particulate Pollution (PM_{10} and $PM_{2.5}$) Forecasting. *Journal of the Air & Waste Management Association* **52**, 1096-1101.

[18] Munir, S., Habeebullah, T., Seroji, A., Morsy, E., Mohammed, A., Saud, W., Abdou, A. and Awad, A. (2013). Modeling Particulate Matter Concentrations in Makkah, Applying a Statistical Modeling Approach. *Aerosol Air Qual. Res.* **13**, 901–910.

[19] Mayers, R. H. (1990). *Classical and modern regression with applications,* 2nd edition, Duxbury Press.

[20] Naes, T. *et al*. (2002). A User-Friendly Guide to Multivariate Calibration and Classification. UK: NIR Publications Chichester.

[21] Ozdemir, U. and Taner, S. (2014). Impacts of Meteorological Factors on PM_{10}: Artificial Neural Networks (ANN) and Multiple Linear Regression (MLR) Approaches. *Environmental Forensics* **15**, 329–336.

[22] Phatak, A. and De Jong, S. (1997). The geometry of partial least squares. *Journal of Chemometrics* **11**, 311–338.

[23] Pires, J. C. M., Martins, F. G., Sousa, S. I. V., Alvim-Ferraz, M. C. M., Pereira, M. C. (2008). Prediction of the Daily Mean PM_{10} Concentrations Using Linear Models. *American Journal of Environmental Sciences* **4:5**, 445-453.

[24] Polat, E. *Partial Least Squares Regression Analysis,* Master Turkish Thesis, Hacettepe University Department of Statistics, Ankara, Turkey, 2009.

[25] Rawlings, J.O. (1988). *Applied Regression Analysis: A Research Tool,* Pacific Grove, California: Wadsworth & Brooks/Cole Advanced Books & Software.

[26] Rosipal R. and Krämer N. (2006). *Overview and Recent Advances in Partial Least Squares,* In: Saunders C, Grobelnik M, Gunn S, Shawe-Taylor J (Eds.), Subspace, Latent Structure and Feature Selection Techniques Springer: 34-51.

[27] Salh, S. M. (2014). Using Ridge Regression model to solving multicollinearity problem. *International Journal of Scientific & Engineering Research* **5:10**, 992-998.

[28] Sayegh, A. S., Munir, S., Habeebullah, T. M. (2014). Comparing the Performance of Statistical Models for Predicting PM_{10} Concentrations. *Aerosol and Air Quality Research* **14**, 653–665.

[29] Syafei, A. D., Fujiwara, A. and Zhang, J. (2015). Prediction Model of Air Pollutant Levels Using Linear Model with Component Analysis. *International Journal of Environmental Science and Development* **6:7**, 519-525.

[30] Tasdemir, Y., Cindoruk, S. S., Esen, F. (2005). Monitorıng of Criteria Air Pollutants in Bursa, Turkey. *Environmental Monitoring and Assessment* **110**, 227–241.

[31] Thomas, E. V. (2003). Non-parametric statistical methods for multivariate calibration model selection and comparison. *Journal of Chemometrics* **17**, 653-659.

[32] Thomas, S. and Jacko, R. B. (2007). Model for Forecasting Expressway Fine Particulate Matter and Carbon Monoxide Concentration: Application of Regression and Neural Network Models. *Journal of the Air & Waste Management Association* **57**, 480-488.

[33] Trygg, J. (2002). Have you ever wondered why PLS sometimes needs more than one component for a single-y vector? *Chemometrics Homepage*, February 2002, http://www.chemometrics.se/editorial/feb2002.html.

[34] Ul-Saufie, A.Z., Yahya A.S., Ramli N.A., Hamid, H. A. (2011a). Comparison Between Multiple Linear Regression And Feed forward Back propagation Neural Network Models For Predicting PM_{10} Concentration Level Based On Gaseous and Meteorological Parameters. *International Journal of Applied Science and Technology* **1:4**, 42-49.

[35] Ul-Saufie, A.Z., Yahya A.S., Ramli N.A. (2011b). Improving multiple linear regression model using principal component analysis for predicting PM_{10} concentration in Seberang Prai, Pulau Pinang. *International Journal of Environmental Sciences* **2:2**, 415-422.

[36] Voukantsis, D., Karatzas, K., Kukkonen, J., Rasanen, T., Karppinen, A. and Kolehmainen, M. (2011). Intercomparison of air quality data using principal component analysis, and forecasting of PM_{10} and $PM_{2.5}$ concentrations using artificial neural networks, in Thessaloniki and Helsinki. *Science of The Total Environment* **409**, 1266-1276.

[37] Wise, B.M. *et al*. (2006). *PLS Toolbox 4.0 for use with Matlab*. 3905 West Eaglerock Drive, Wenatchee, WA, Eigenvector Research Inc. http://www.eigenvector.com

[38] Wold, S. *et al*. (2001). PLS-regression: a basic tool of chemometrics. *Chemometrics and Intelligent Laboratory Systems* **58**, 109-130.

[39] Yusof, N. F., Ghazali, N. A., Ramli, N.A., Yahaya, A. S., Sansuddin, N., Al Madhoun, W.A. (2008). Correlation of PM_{10} Concentration and Weather Parameters in Conjunction with Haze Event in Seberang Perai, Penang. *Proceeding of The International Conference on Construction and Building Technology 2008 (ICCBT08) in Kuala Lumpur, Malaysia*.

Received March 15, 2013; accepted November 10, 2013.

Esra Polat
Department of Statistics, Faculty of Science
Hacettepe University, 06800, Ankara
Turkey Faculty of Applied Sciences
espolat@hacettepe.edu.tr