International Journal of Applied Mathematics
& Statistical Sciences (IJAMSS)
ISSN(P): 2319-3972; ISSN(E): 2319-3980
Vol. 7, Issue 1, Dec - Jan 2018; 1 - 12

International Academy of Science, Engineering and Technology
Connecting Researchers; Nurturing Innovations

A PRINCIPAL COMPONENT REGRESSION MODEL, FOR FORECASTING DAILY PEAK AMBIENT GROUND LEVEL OZONE CONCENTRATIONS, IN THE PRESENCE OF MULTICOLLINEARITY AMONGST PRECURSOR AIR POLLUTANTS AND LOCAL METEOROLOGICAL CONDITIONS: A CASE STUDY OF MAUN

W. M. Thupeng[1], T. Mothupi[2], B. Mokgweetsi[3], B. Mashabe[4] & T. Sediadie[5]

[1,2,3,5]Department of Statistics, University of Botswana, Gaborone, Botswana

[4]Department of Mathematics and Statistical Sciences, Botswana International University of Science and Technology, Palapye, Botswana

ABSTRACT

The increasing public awareness of the negative health effects of exposure to peak air pollution levels, particularly to the most sensitive population sub-groups like children and the elderly, has made short-term forecasts of episodes of peak concentrations of air pollutants at a local level, increasingly necessary. The main objective of the present study is to develop a statistical model for predicting a day in advance the daily maximum 1- hour average ambient ground level ozone concentration for Maun town, using principal component regression (PCR) technique. The predictor variables are the precursor air pollutants of ground level ozone, namely, nitrogen oxides, nitrogen dioxide and the previous day's ground level ozone concentration, on the one hand, and meteorological variables that include wind speed, wind direction, relative humidity, surface temperature, atmospheric pressure and solar radiation. The data consist of maximum 1-hour interval concentrations every day, on the response and each of these predictor variables collected from 1 May 2014 to 30 September 2015. A biased regression method of PCR is applied to try and minimise the problem of multicollinearity, usually associated with multiple regression models. The detection of multicollinearity is performed by using the Pearson partial correlation matrix, and variance inflation factor (VIF). Model assessment tools include the tests for significance of individual regression coefficients in the PCR model, the coefficient of determination and F test to test for the validity of the overall model. It is found that the estimated PCR model is based on principal components that are highly correlated with maxima of the ozone concentration the day before, nitrogen oxides concentrations and surface temperature. Furthermore, wind speed, wind direction, relative humidity and nitrogen dioxide are identified as possible causes of multicollinearity, in the available data.

***KEYWORDS**: Ambient Air Pollution, Meteorological Conditions, Multicollinearity, Peak Ambient Ground level Ozone, Precursor Air Pollutants, Principal Component Regression*

Article History

***Received:** 22 Nov 2017 | **Revised:** 12 Dec 2017 | **Accepted:** 19 Dec 2017*

1. INTRODUCTION

Environment-related problems such as water and air pollution have attracted much greater research attention, in the twenty-first century, than ever before. Ambient (outdoor) air pollution is a major environmental health problem affecting everyone in developed and developing countries, a like [1]. In particular, as pointed out by [2], the problem of air pollution in cities has become so severe that, there is a need for timely information about changes in the pollution level. Nowadays, ozone is considered as one of the most significant air pollutants, owing to the fact that, it severely affects plant tissues and human health [3].

According to [4], NOx and VOCs (especially non-methane hydrocarbons, NMHC) and carbon monoxide (CO) are among the most important O_3 precursors. Biomass burning is also a major source of trace gases and particulates [5]. According to [5], in the CAPIA project, a threshold value of 40 ppb is used to assess the potential risk of damage to maize by ozone and, measured data show that this threshold is exceeded over Botswana and on the South African Highveld. Also, [6] report that in Botswana, monitoring is ongoing at Maun where concentrations of 90 ppb and higher are not uncommon. These conditions, coupled with meteorological conditions that may interfere with the dispersion of ozone in the atmosphere or result in increased production of pollutants, are conducive to the formation of ozone, suggesting that ozone concentration over Botswana, particularly Maun, may be relatively high.

The influence of local climatic factors on groundlevel ozone concentrations is an area of increasing interest to air quality management in regards to future climate change [7]. Several methods have been used for modelling ambient groundlevelO_3 concentrations. Among these methods, multiple linear regressions have provided successful results in modelling studies [8]. The key assumption of a multiple regression model is that the predictor variables are independent. However, multiple linear regression models usually have a problem of multicollinearity (collinearity in the predictors), a condition that is due to high correlations amongst some of the predictors in the model. The adverse effect of multicollinearity is that the resulting ordinary least squares (OLS) estimates of the regression coefficients of the correlated predictor variables tend to have large sampling variances, which makes the estimates unstable. Thus, multicollinearity leads to spurious results and, consequently, wrong conclusions of a study.

To try and minimise the problem of multicollinearity, biased regression methods are applied. A biased regression method stabilises the partial regression coefficients of a model by introducing bias. The bias is associated with a reduction in the variance of the estimated coefficients, so there is a gain that more than compensates for the increase in bias [9]. In the statistical literature, amongst the most common biased regression methods like PCR, ridge regression (RR) and partial least squares (PLS) regression, the PCR appears to be the most widely employed in the atmospheric sciences. For instance, [10] compare multiple linear regression, feed forward artificial neural networks using principal components as inputs and, principal component regression to predict next day hourly ozone concentrations using as predictors air pollutant concentrations of NO, NOx, on the one hand, and hourly means of surface temperature, wind velocity and relative humidity. [4] Apply both multiple linear regression and principal component regression techniques to predict concentrations at the ground level of the troposphere as a function of several air pollutants and meteorological parameters. More recently, [11] compare multiple regression and principal regression techniques to forecast total column of ozone with other 8 ambient atmospheric parameters as predictor variables over Peninsular Malaysia. Most recently, [12] use multiple regression analysis and principal component analysis techniques to develop models for the prediction of column ozone concentrations with a few selected atmospheric parameters as predictors for Peninsular Malaysia.

Identifying an appropriate probability model to describe the stochastic behaviour of extreme ambient air pollution level for a specific site or multiple sites forms an integral part of environmental management and pollution control [13]. In Botswana, monitoring of ambient air groundlevel ozone by DWMPC is ongoing at a number of monitoring sites. However, the available literature does not show if there has been any systematic study on the determination of suitable statistical models for modeling concentrations of ambient groundlevel ozone in any of the monitoring stations in Botswana. Therefore, the present study attempts to model one day in advance the daily maximum 1-hour average ambient groundlevel ozone concentrations in the presence of precursor pollutants and local meteorological conditionsfor Maun.

2. DATA AND RESEARCH METHODOLOGY

2.1. Site Description

Maun is a tourist centre in Botswana, being the gateway to the pristine Okavango Delta. So, the managers of air quality in the areas are under pressure from government, citizen, and other local businesses to maintain a high level of air quality for the health benefits of the community and to make the area attractive to tourists and new business and industry. The town has shopping centres, hotels and lodges as well as car hire, and the busiest airport in Botswana. This monitoring site is located in the high motor vehicle traffic near the main bus, taxi station and the airport, where the amount of nitrogen oxides emitted into the atmosphere as air pollution can be quite high. The station is selected because it consistently indicated very high levels of concentrations of ground level ozone. Furthermore, this particular weather station is chosen based on availability, reliability and good quality of the data. In particular, the data availability for the site for the months of May through August, 2014, was 100%. However, data for 2014 or the past years were not available.

2.2. Data Description

The data used for this study consist of 123 the daily maximum 1-hour average ambient groundlevel ozone concentrations of O_3, 3 precursor pollutants of groundlevel ozone, nitrogen oxides (NOx), nitrogen dioxide (NO_2), the day before' s peak groundlevel ozone concentration (O_{t-1}), all measured in parts per billion (ppb) by volume, and 8 meteorological variables, namely, wind speed (WS), measured in meters per second, wind direction (WD), in degrees, relative humidity (RH), in percentage, surface temperature (T) in degrees Celsius, atmospheric pressure (P), measured in millibars and solar radiation (R), measured in Watts per square meter. These data were obtained from DWMPC, Ministry of Environment Wildlife and Tourism, Botswana. Datasets from the DWMPC are considered to be of high quality, with no long periods of missing values over the study period.

2.3. Principal Component Regression Model

For day $t, t = 1, 2, 3, ..., 123$, denote by Y_t the maximum groundlevel ozone concentration, with the predictor variables$O_{3(t-1)}$, WS,WD,, T, RH, P, R, NOx and NO_2. as defined before.Principal components regression is a regression technique that is based on principal component analysis (PCA). Without loss of generality, let us denote the original predictor variables by $X_1, X_2, ..., X_p$. PCA is a multivariate technique that seeks to explain the variance-covariance structure of the p original variables, $X_1, X_2, ..., X_p$, through a few uncorrelated linear combinations, $Z_1, Z_2, ..., Z_p$, of the original variables. The basic idea behind PCR is to calculate the principal components and then use some of them as predictors in a linear regression model fitted using the ordinary least squares (OLS) method. PCR can minimise

multicollinearity by excluding some of the low-variance principal components in the regression model.

To construct the PCR model in matrix form, we start with the usual multiple linear regression model given by Equation (1).

$$\mathbf{y}_{(nx1)} = \mathbf{X}_{(nxp)}\boldsymbol{\beta}_{(px1)} + \mathbf{e}_{(nx1)} \quad \textbf{(1)}$$

where n is the sample size, $\mathbf{y}$ is a vector of values of the dependent variable, $\mathbf{X}$ is a matrix of predictor variables, which has a multivariate distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$, $\boldsymbol{\beta}$ is a vector of regression coefficients and $\mathbf{e}$ is the error term. The key assumptions of the model are (i) a linear relationship between the response variable and the predictor variables, (ii) the independent variables are not highly correlated with each other and (iii) the errors are jointly normally distributed independent variables with a mean of zero and constant variance. The OLS estimator of $\boldsymbol{\beta}$ is given by $\hat{\boldsymbol{\beta}} = \left(\mathbf{X'X}\right)^{-1}\mathbf{X'Y}$. It must be noted that the variables considered in the present study are expressed in different units of measurement. To correct for the varying units of the variables and also allow for the independent variables to be treated as equally important, all the independent variables are scaled/standardised by subtracting the mean of each variable from the variable and then dividing by the corresponding standard deviation, so that that all the variables have mean zero and unit variance. and, as a result of standardisation $\mathbf{X}'\mathbf{X} = \mathbf{R}$. Effectively, this means that the principal component analysis is performed by using the correlation matrix **R** of the original data.Thus, to obtain the PCR model, we transform the original variables, $X_1, X_2, ..., X_p$, to the principal components , $Z_1, Z_2, ..., Z_p$, using Equation (2).

$$\mathbf{X}'_{(pxn)}\mathbf{X}_{(nxp)} = \mathbf{P}_{(pxp)}\mathbf{D}_{(pxp)}\mathbf{P}'_{(pxp)} = \mathbf{Z}'_{(pxp)}\mathbf{Z}_{(pxp)} \quad \textbf{(2)}$$

Where $\mathbf{D}$ is the diagonal matrix of the eigenvalues of $\mathbf{X}'\mathbf{X}$, $\mathbf{P}$ is the eigenvector matrix of $\mathbf{X}'\mathbf{X}$ (and is orthogonal as $\mathbf{P}'\mathbf{P} = I$) and, $\mathbf{Z}$ is a matrix of the principal components $Z_1, Z_2, ..., Z_p$, with successively smaller variances so that $\text{var}(Z_1) \geq \text{var}(Z_2) \geq, ..., \text{var}(Z_p)$, where $z_j = \beta_{1j}x_1 + \beta_{2j}x_2 + ... + \beta_{pj}x_p, j = 1, 2, ..., p$. Therefore, the PCR model to be used for prediction of the maximum 1-houraverage groundlevel ozone concentration a day in advance, Yt , is given by Equation (3).

$$y_t = \alpha_1 z_{1t} + \alpha_2 z_{2t} + ... + \alpha_p z_{pt}, t = 1, 2, ..., n \quad \textbf{(3)}$$

where z_{jt} denotes the score on the j[th] principal component on day t, $j = 1, 2, 3, ..., p$ and $t = 1, 2, ..., n$.

2.4. Detection of Multicollinearity

2.4.1. Examination of Partial Correlation Coefficient

To examine relationships between environmental variables, which are inherently highly correlated, it is better to use pair wise partial correlation coefficients rather than ordinary correlation coefficients, which indicate how much each variable uniquely contributes to the coefficient of determination, R^2 , over and above that which can be accounted for by the other predictors**.** The partial correlation coefficient between variables X_i and X_j, $i \neq j = 1, 2, ..., p$, holding all the other

variables constant is given Equation (4).

$$r_{x_i,x_j.z} = -\frac{r_{ij}}{\sqrt{r_{ii}r_{jj}}} \quad \textbf{(4)}$$

where $r_{ij} = r_{ji}, i \neq j = 1, 2, ..., p$, is the $(i,j)^{th}$ element of the inverse of the correlation matrix $\boldsymbol{R}$ of $\mathbf{X}$,$\boldsymbol{R}^{-1} = [p_{ij}]$. Generally, a (partial) correlation coefficient of at least 0.80 is an indication that multicollinearity may exist.

2.4.2. Variance Inflation Factor

The VIF is a measure of how much the variance of the estimate of the ordinary least squares (OLS) regression coefficient is being inflated by multicollinearity. Given a set of p predictor variables, $X_1, X_2, ..., X_p$, the VIF for the jth variable is obtained by regressing X_j on all the remaining predictor variables in the model, and is given by Equation (5).

$$VIF_j = \frac{1}{1-R_j}, j = 1, 2, ..., p \quad \textbf{(5)}$$

Where R_j denotes the coefficient of determination of the regression model of X_j, on all the other predictors. A value of $VIF_j \geq 10$ (or tolerance (=1/VIF) value less than 0.1) roughly indicates significant multicollinearity.

2.4.3. Eigenvalues of the Correlation Matrix

The eigenvalues of **R,** denoted by $\lambda_1, \lambda_2, ..., \lambda_p$, and usually ordered in decreasing magnitude, can be used to study the correlation structure of **R,** whose full rank is given by the total number of the eigenvalues. When there is no multicollinearity, the eigenvalues $\lambda_j, j = 1, 2, ..., p$, will all be equal to 1. A zero λ_j indicates linear dependency and an eigenvalue close to zero indicates a near linear dependency [14].

2.4.4. Determining the Number of Components

One of the tools for determining the number of principal components to be used for further analysis is to examine the cumulative proportion to determine the amount of variance that the principal components explain. Retain the principal components that explain an acceptable level of variance, depends on the application. For example, if we want to perform further analyses on the data, we may want to have at least 90% of the variance explained by the principal components. We can also use eigenanalysis method, which uses the size of the eigenvalue to determine if its associated principal component should be included for further analyses. Using the Kaiser criterion, we use only the principal components with eigenvalues that are greater than 1.

A widely used tool for deciding on the number of principal components to retain in a PCA is the scree plot obtained by plotting ordered pairs of values$(j, \lambda_j), j = 1, 2, \ldots, p$. The basic rule is to select the components in the steep curve before the first point that starts the line trend.

2.4.5 Assessing Model Adequacy

To test whether there is sufficient evidence of a linear relationship of an individual predictor and the response variable, we test the hypothesis $H_0: \beta_j = 0$ against the alternative $H_0: \beta_j \neq 0$ for $j = 1, 2, \ldots, p$.

UnderH_0, the student's t statistic is defined by Equation (6)

$$t = \frac{b_j}{se(b_j)} \sim t_{n-p-1} \quad \textbf{(6)}$$

and, the null hypothesis is rejected in favour of the alternative if the calculated t value is improbably large.

The F-test is used to check for the validity of the overall model. That is, $H_0: \beta_1 = \beta_2 = \cdots = \beta_p = 0$ against the alternativeH_0: At least one$\beta_j \neq 0$ for $j = 1, 2, \ldots, p$. The test statistic is given by

$$F = \frac{MSR}{MSE} \sim F_{(p;\, n-p-1} \quad \textbf{(7)}$$

Where MSE and MSR denote the mean square for error and mean square for regression, respectively. The decision rule is to reject the null hypothesis in favour of the alternative if the calculated F value in (7) is improbably large. The overall model is further validated by considering the Coefficient of determination (R^2) defined by

$$R^2 = 1 - \left[\frac{\sum_{t=1}^{n} \hat{\varepsilon}_t^{\,2}}{\sum_{t=1}^{n} (Y_t - \overline{Y}_t)^2} \right] = \left[\frac{\sum_{t=1}^{n} (\hat{Y}_t - \overline{Y}_t)^2}{\sum_{t=1}^{n} (Y_t - \overline{Y}_t)^2} \right] \quad \textbf{(8a)}$$

Where Y_t is the observed ozone concentration at time t, $\hat{Y}_t$ is the predicted ozone concentration at time t and n is the number of observations. The coefficient of determination R^2 in Equation (8a), and the adjusted r-squared $R_{adj}^{\,2}$ give in Equation (8b), give the proportion of the total variability in the response Y explained by the fitted regression model. The adjusted R-squared provides an adjustment for the degrees of freedom.

$$R_{adj}^{\,2} = 1 - \left[\frac{(1 - R^2)(n-1)}{n-k-1} \right] \quad \textbf{(8b)}$$

Where n = the number of sample observations, R^2 =coefficient of determination and k =number of independent repressors.

3. RESULTS AND DISCUSSIONS

3.1 Time Series Plot of the Data

Figure 1 shows a time series plot of the daily maximum 1-hour average ambient groundlevel ozone concentrations for Maun for the study period. From the figure, most important to notice is that during August 2014, the threshold value of 40 ppb used in the CAPIA project to assess the potential risk of damage to maize by

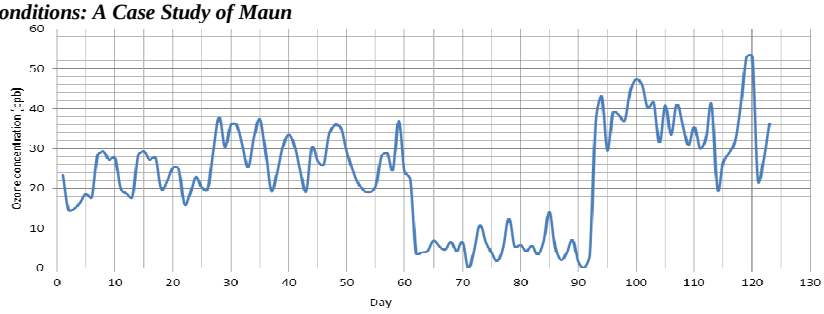

Figure1: Time series plot of daily maximum groundlevel ozone concentrations for Maun, 1 May - 31 August, 2014

Figure 1

Ozone was exceeded over Maun about 10 times. The highest daily 1-hour average O_3 concentration of 53.079 ppb occurred on the 28th of August 2014 (day 120). Also, the month of August experienced most variability in the daily peak ozone concentrations.

3.2. Detecting Multicollinearity

3.2.1. Examination of Pairwise Partial Correlation Matrix

From the matrix of pairwise correlations for predictor variables (not presented in the paper), it can be seen that some of the predictor variables are highly correlated even after controlling for the effects of the other predictors. For example, when the other predictor variables are kept constant, the correlation between WD and WS is positive and very high ($r_{23.x} = 0.945$). A similar observation can be made about RH and NO_2, with $r_{58.x} = 0.802$. Also noteworthy is that the partial correlation between NO_2 and NO_X is fairly strong ($r_{89.x} = 0.631$) .This is to be expected as nitrogen oxides (NOx) in the ambient air consist primarily of nitric acid (NO) and nitrogen dioxide (NO_2) (i.e., NOx=NO+NO_2). These results show that some of the predictor variables in this study are highly correlated, which is a sign of the possibility of the existence the problem of multicollinearity.

3.2.2. Variance Inflation Factor

Results in Table 1 clearly show that the VIF values for the variables WS, WD, RH and NO2 are each larger than the threshold value of 10, indicating that a high degree of multicollinearity exists and may be due to these variables. These results confirm those of the partial correlation analysis and the VIF that the problem of multicollinearity amongst some predictor variables in this dataset exists. Therefore, to control and minimize the effects of multicollinearity, we model the Maun daily maximum ground level ozone data using the principal component regression technique.

3.3. Principal Component Regression

For these reasons stated earlier on, the covariance matrix is scaled so that the principal component analysis is performed by using the correlation matrix. Table 4.1 shows the principal components computed from the correlation matrix.

3.3.1. Determining the Number of Components

The matrix of weights of the principal components computed from the correlation matrix has not been provided in this paper for a lack of space. However, it must be noted that the magnitudes of the coefficients also dependents on the variances of the corresponding variables. and, to achieve a simple structure that makes interpretation of the principal components as intuitively as possible, a rotation of the axes (dimensions) of the first few selected principal components is carried out using the varimax rotation, a popular orthogonal rotation method.

We discuss the results of the varimax rotation (contained in Table 3) of the selected five principal components in the sequel before fitting the PCR model.

To determine the minimum number of principal components that account for most of the variation in the data, we use the eigenanalysis of the correlation matrix and the Scree plot, shown in Table 2 and Figure 2, respectively. In Table 2, the eigenvalues of the principal components show that the data has the largest variance along the component 1 axis and the second largest variance along the axis of component 2, and so on. Principal component z_1 alone contributes 67.3% to the total variance in the original variables, followed by the 2nd highest of 22.2% by z_2. The principal components z_1 through z_6 have eigenvalues $\lambda_j \geq 1$. Thus, using the Kaiser criterion, we would use only the principal components with eigenvalues that are greater than 1, the first 6 principal components. However, $z_6$through z_9 contribute increasingly very little or nothing to the total variance. and, cumulatively, the first 5 components contribute 98.9% to the variability in the original variables. Hence, on balance, only the first 5 principal components should be retained for further investigation.

To further help in the choice of the number of principal components to be selected for inclusion in the model, the scree plot for the data on the predictor variables in the present study is given in Figure 2. Noticeably, an elbow occurs at dimension j=3 in the scree plot but, the slope of the graph becomes fairly constant after λ_5 and all the following eigenvalues become relatively small and about the same size. Therefore, it is decided that the first 5 principal components, z_1 through z_5, should effectively explain the total variance in the original variables. Consequently, we fit a principal regression model of O3 on the principal components,z_1, .

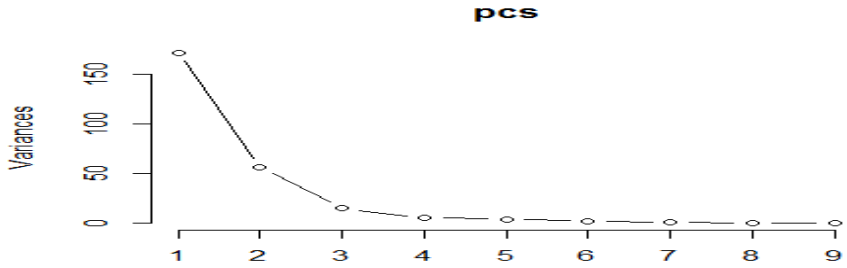

Figure 2: Screen Plot of Precursor Ambient Air Pollutants and Meteorological Variables for Maun

As mentioned before, to aid in the interpretation of principal components, we performed a varimax rotation of the selected first five principal components. The rationale for the varimax rotation is to produce axes with a few large loadings and as many near-zero loadings as possible. Loadings are regarded as correlation coefficients between the original variables and their respective principal components, with the loading for principal component z_j given by $l_j = e_j\sqrt{\lambda_j}, j = 1,2,\ldots,p$. Inthe present study, the loadings are expected to be very similar to the principal component scores since all the variables have equal (unit) variance. The results in Table 3 show that the first principal z_1 correlates almost perfectly with $O_{3(t-1)}$.That is, it increases with the previous day's peak groundlevel ozone concentration. So, this component mainly measures background groundlevel ozone concentration at the monitoring station This result seems to indicate that groundlevel ozone concentrations are persistent so that the current level of concentrations are most likely to build on the previous day's peak groundlevel ozone concentration. The second principal component increases with decreasing concentrations of nitrogen dioxide NO_2 and, to a less extent, relative humidity,

RH. This means the second component z_2 primarily measures the concentration of ambient air nitrogen dioxide and relative humidity at the monitoring site. Thus, these two air pollutants vary together. Similarly, the third component is highly correlated with only one predictor variable, NOx, indicating that z_3 increases with nitrogen oxides. So, z_3 primarily measures ambient air concentration of oxides of nitrogen at the station. Likewise, components z_4 and z_5 mainly measure wind direction and surface air surface temperature, respectively.

3.3.2. Fitting the Principal Component Regression Model

A summary of the results of fitting the principal component regression model of O_3 on the principal components, z_1, z_2, z_3, z_4 and z_5 are contained in Table 4. The large value of F (94.7) with a p-value: < 0.001 is enough evidence that the overall model is valid. In addition, $R^2 = 0.8019$ and $R^2_{adj} = 0.7934$ both show that a high proportion (about 80%) of the variability in the concentration of ambient air groundlevel ozone in Maun is explained by variability in the componentsz_1, z_3 andz_5. However, the t test for significance of the individual components suggests that z_2 and z_5 individually is not significant. This means that their contribution in the formation of O3 is not important and, therefore, they can be excluded from final the model. Clearly, the results suggest that there are three important components for forecasting a day in advance the daily maximum 1-hour average groundlevel ozone for Maun components are z_1, z_3 andz_5, all of which are highly significant with a p-value: < 0.001 each. Therefore, the estimated PCR model for forecasting one day in advance the daily maximum 1-hour average groundlevel ozone for Maun is given by Equation (9) as

$$O_3 = 0.83297z_1 + 0.90783z_3 + 0.91919z_5 \quad \textbf{(9)}$$

4. CONCLUSIONS

This paper presents a principal regression model for forecasting a day in advance the daily maximum 1-hour average groundlevel ozone for Maun using precursor ambient air pollutants for ozone and local meteorological conditions as predictor variables. The model minimises the effects of multicollinearity amongst the predictors, which is inherent in environmental data, by excluding some of the low-variance principal components. It is found that estimated PCR model is based on principal components that are highly correlated with three predictor variables: the day before's groundlevel ozone concentration, concentrations of nitrogen oxides and surface temperature. The estimated PCR equation is easy to implement and can be adapted to other air pollution monitoring stations around Botswana.

There are two limitations of the study. First, data on other important variables contributing to groundlevel ozone formation, like VOCS and carbon monoxide are currently not measured at the station. Secondly, data for the summer months were not available, at least at the time of collecting the data for the study, and the available data were for only 4 months. However, the results of this study have highlighted the important relationship between groundlevel ozone concentrations and local ambient air pollution levels and meteorological conditions in the study area.

Table 1: Individual Multicollinearity Diagnostics for Air Pollutants and Meteorological Parameters

Variable	$O3_{(t-1)}$	WS	WD	T	RH	P	R	NO_2	NO_x
VIF	1.8336	17.4160	15.2253	1.5152	17.6478	1.9463	6.0510	13.6290	1.9826

Table 2: Matrix of Weights of the Principal Components Computed from the Correlation Matrix

Component	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9
Eigenvalues λ_j	171.607	56.620	14.750	5.304	3.903	1.866	0.779	0.090	0.010
Proportion of Variance	0.673	0.222	0.058	0.021	0.015	0.007	0.003	0.000	0.000
Cumulative Proportion	0.673	0.895	0.953	0.974	0.989	0.997	1.000	1.000	1.000

Table 3: Varimax Rotation

	z_1	z_2	z_3	z_4	z_5
O3.t.1.	0.999	-	-		-
WS	-	-	-	-0.154	-
WD	-	-	-	-0.946	-
T	-	-	-	0.123	-0.930
RH	-	-0.566	-0.176	-	0.121
P	-	-	0.118	0.248	0.325
R	-	-	-	-	-
NO2	-	-0.812	0.253	-	-0.107
NOx	-	0.105	0.942	-	-

Table 4: Estimates of the Air Pollutants and Meteorological Parameters Based on the PCR Model

Variable	Estimate	Std. Error	t value	P-value
Intercept	-166.68938	95.15875	-1.752	0.082
z_1	0.83297	0.04075	20.442	< 0.001 ***
z_2	-0.08170	0.07094	-1.152	0.252
z_3	0.90783	0.13899	6.532	< 0.001 ***
z_4	0.05694	0.23179	0.246	0.806
z_5	0.91919	0.27019	3.402	< 0.001 ***
Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1				
Residual standard error: 5.896 on 117 degrees of freedom, Multiple R-squared: 0.8019, Adjusted R-squared: 0.7934				
F-statistic: 94.7 on 5 and 117 DF, p-value: < 0.001				

REFERENCES

1. *WHO (World Health Organization) Media Centre. "Ambient (outdoor) air quality and health". Fact Sheet. UpdatedSeptembers 2016. http://www.who.int/mediacentre/factsheets/fs313/en/. Accessed 21 August 2017.*
2. *Kumar, A. and Goyal, P. (2011). Forecasting of air quality in Delhi using principal component regression technique. Atmospheric Pollution Research* ***2****, 436 – 444.*
3. *Pavon-Dominguez, P., Jimenez-Hornero, F.J., and Gutierrez de Rave, E. (2014). Proposal for estimating groundlevel ozone concentrations at urban areas based on multivariate statistical methods. Atmospheric Environment* ***90****, 59-70.*
4. *Ozbay, B., Keskin, G.A., Dogruparrmak, S.C. and Ayberk, S. (2011). Multivariate methods for groundlevel ozone modelling. Atmospheric Research* ***102*** *(2011), 57–65.*

5. *Zunckel, M., Venjonoka, K., Pienaar, J.J., Brunke, E-G., Pretorious, O., Koosialee, A., Raghunandan, A. and van Tienhoven, A.M. (2004). Surface ozone over southern Africa: synthesis of monitoring results during the Cross border Air Pollution Impact Assessment project. AE International – Africa & the Middle East. Atmospheric Environment* ***38****, 6139-6147.*

6. *Zunckel, M., Koosialee, A., Yawood, G., Maurer, G., Venjonoka, K., van Tienhoven, A.M. and Otter, L. (2006).Modelled surface ozone over southern Africa during the Cross Border Air Pollution Impact Assessment Project. Environmental Modelling and Software* ***21****, 911-924.*

7. *Jalaludin J, Nordiyana M. S & Suhaimi N. F, Exposure to Indoor Air Pollutants (Formaldehyde, VOCS, Ultrafine Particles) and Respiratory Health Symptoms among Office Workers in Old and New Buildings in University Putra Malaysia, International Journal of Applied and Natural Sciences (IJANS), Volume 3, Issue 1, December-January 2014, pp. 69-80*

8. *Gorai, A.K., Tuluri, F., Tchounwou, P.B. and Ambinakudige, S. (2015). Influence of local meteorology and NO_xconditions on groundlevel ozone concentrations in the eastern part of Texas, USA. Air Quality, Atmosphere and Health* ***8****, 81-96.*

9. *Abdul-Wahab, S.A., Bakheitb, C.S., Al-Alawi, S.M., (2005). Principal component and multiple regression analysis in modelling of groundlevel ozone and factors affecting its concentrations. Environmental Modelling &Software****20****, 263–1271.*

10. *S. Christy & V. Khanaa, The Effects of Air Pollution on Human Health, International Journal of Mathematics and Computer Applications Research (IJMCAR), Volume 6, Issue 1, January-February 2016, pp. 51-58*

11. *University of California Davis (206). Lecture outline AGR 206 Chapter 9. Biased Regression. AGR206Ch09PCR.doc. www.plantsciences.ucdavis.edu/agr206/agr206files/.../agr206ch09pcr.pdf.*

12. *Sousa, S.I.V., Martins, F.G., Alvim-Ferraz, M.C.M., Pereira, M.C. (2007). Multiple linear regression and artificial neural networks based on principal components to predict ozone concentrations. Environmental Modelling &Software****22****, 97–103.*

13. *Rajaba, J.M., MatJafrib, M.Z. and Limb, H.S. (2013). Combining multiple regression and principal component analysis for accurate predictions for column ozone in Peninsular Malaysia. Atmospheric Environment, vol* ***71****, 36-43.*

14. *Muhammad Roman & Muhammad Idrees, A Qualitative Study of Causes and Effects of Air Pollution on Human Health in Faisalabad Pakistan, International Journal of Environment, Ecology, Family and Urban Studies (IJEEFUS), Volume 3, Issue 1, March-April 2013, pp. 139-146*

15. *Tan, K.C., Lim, H.S., Mat Jafri, M.Z (2016). Prediction of column ozone concentrations using multiple regression analysis and principal component analysis techniques: A case study in peninsular Malaysia. Atmospheric Pollution Research* ***7****, 533-546.*

16. *Thupeng, W.M. (2016). Use of the Three-parameter Burr XII Distribution for Modelling Ambient Daily Maximum Nitrogen Dioxide Concentrations in the Gaborone Fire Brigade. American Scientific Research Journal for Engineering, Technology, and Sciences (ASRJETS), vol* ***26****, No. 2, 18-32.*

17. *Jobson, J.D. (1991). Applied multivariate Data Analysis. Volume I: Regression and Experimental Design. Springer Science & Business Media New York.*