Volume34 Number 1
Mar 2023

Journal of Energy in Southern Africa

Modelling emissions from Eskom's coal fired power stations using Generalised Linear Models

D. Chikobvu[1]*, M. Mamba[2]

[1]Department of Mathematical Statistics and Actuarial Science, University of the Free State, Mangaung, South Africa
[2]Department of Mathematical and Physical Sciences, Central University of Technology, Mangaung, South Africa

Abstract

The aim of this paper is to determine if a Generalised Linear Model (GLM) is a better model over the traditional simple linear regression when fitted to nitrogen dioxide (NO2) emitted into the atmosphere during the production of electricity from 13 Eskoms coal fuelled power stations. GLMs have flexibilities of allowing the variance to vary as a function of the mean (non-constant variance), and have the advantage of keeping the data in its original scale. Unlike regression, the models do not assume a linear relationship between the response variable and the explanatory variables, and instead the link function is used. The data also need not be Normally distributed. Group-lasso interaction network (glinternet) was used in variable selection for the GLM models. A similar model using regression analysis was fitted for comparison. The results show that a GLM can be used to predict and explain NO_2 emissions from coal fired electricity stations in South Africa. The Lognormal model was found to be the better model by diagnostic measures including plots that showed improved variance behavior in the residuals. Various variables such as amount of electricity sent out (in GWhs), age of power station (in years), power station used, and interaction terms such as electricity and station, Age and station can be used in describing and predicting NO_2 emissions (in tons) from Eskoms coal fuelled power stations.

Keywords: *Eskom; generalised linear model(s) (GLM); linear regression; lognormal distribution; nitrogen dioxide (NO_2) emissions.*

Journal of Energy in Southern Africa 34(1): 1-14
DOI: https://dx.doi.org/ 10.17159/2413-3051/2022/v33i4a13819
Published by the University of Cape Town ISSN: 2413-3051 https://energyjournal.africa/

Sponsored by the Department of Science and Innovation

*Corresponding author: Email: Chikobvu@ufs.ac.za

Introduction

Coal is the primary source of energy in South Africa and its use has increased significantly over the years. This is as a result of an increased demand of electricity in South Africa. This has given rise to more emission of pollutants including nitrogen dioxide (NO_2) from the coal fired electricity power stations (Eskom, 2016). Exposure to this emission impacts on human health (Anand, Varma and Srimurali, 2013; World Health Organization, 2013; Wellenius, Schwartz and Mittleman, 2015). In order to control NO_2, and sulphur dioxide (SO_2) emissions and other pollutants from the electricity industry, minimum emission standards were published in terms of the National Environmental Management: Air Quality Act in 2010, requiring Eskom to install many retrofits of abatement technologies in order to comply with the emission standards (Eskom, 2011).

The selection of the right statistical probability distribution for describing or modelling environmental pollution data is an important step. These probability models have become the basis for quantifying emissions to meet the evolving information needs of environmental quality management (Singh et al., 2001).

Georgopoulos and Seinfeld (1982) concluded that air pollutant concentrations are inherently random variables because of their dependence on the fluctuations of a variety of meteorological and emission variables. They also concluded that there is no single statistical distribution which gives the best fit to air quality/emission at all time periods. The choice of a statistical distribution generally depends on the pollutant, the time period of interest, the average time of the data, the location and other factors.

Popular statistical probability density functions in representing atmospheric concentrations emissions include the two-parameter distributions (namely, the Lognormal, the Weibull and the Gamma), three-parameter distributions (namely, the 3 parameter Lognormal, the 3 parameter Gamma, the 3 parameter Weibull and 3 parameter Beta distributions) and four-parameter distributions (e.g. four parameter Beta distribution) (Georgopoulos and Seinfeld, 1982). The distributions are useful because of their property of being right skewed, allowing for the modelling of higher emissions.

Statement of the problem

The aim of this study is to determine if Generalised Linear Models (GLMs) have an advantage or give a better model fit than the traditional linear regression model when fitted to the NO_2 emission data. The study also aims to determine those variables contributing significantly to the amount of NO_2 emitted into the atmosphere during the production of electricity from 13 Eskoms coal fuelled power stations.

Justification of the study

The identification of input variables that contribute to the NO_2 emission is important to combat and monitor high emission volumes into the atmosphere in order to find ways to decrease such emissions and meet statutory regulations and lower the risk associated with electricity production emissions. The flexibilities of GLMs, compared to models based on regression analysis, can be useful in the determination of these input variables. This flexibility includes advantages of allowing the variance to vary as a function of the mean (non-constant variance), and the response variables having a distribution other than the Normal distribution. Also, GLMs provide the advantage of keeping the data to its original scale by making use of link functions. In the South African context, there is not sufficient literature to suggest a wide use of GLMs in the modelling of emission, especially the NO_2 pollutant. The study will try to reduce this gap.

Objectives of the study

In this study, the objectives are:

To check if the Lognormal distribution based GLM is a better model over the traditional simple linear regression (the Normal distribution based GLM with identity link function) when fitted on the response NO_2 emission data.

Determining if the variables electricity sent out (GWhs), age of power station (years), power station, abatement technology and month can be used to predict the emission of NO_2 (tons).

To rank the Eskom power plants in terms of NO_2 emission efficiency.

Contribution of the study

With the aging of the power stations and high demand of electricity, NO_2 emissions are projected to increase from coal fired electricity stations in South Africa (Pretorius et al., 2015). There is therefore a need to model NO_2 emissions from these stations. This will provide information to monitor and manage emissions to meet the regulations and thus minimise the exposure of high emissions to humans and the environment.

The rest of the paper is organised as follows, section 2 gives the Literature review. Section 3 gives the methodology. Section 4 gives the results and section 5 concludes.

Literature review

This section reviews some of the literature, including models used in modelling emissions.

Perez and Trier (2001) used predictions to compare linear regression and multilayer neural networks to find a method of predicting NO and NO_2 concentrations. A feed forward neural network was chosen as the convenient method of prediction over the linear regression since this method had reasonable control over the adjustment of parameters.

In studies by Nagendra and Khare (2006), Perez and Trier (2001) and many others, such as those by

Kukkonen et al. (2003) and Capilla (2014), there was a strong non-linear dependency between NO_2 emissions (concentrations) and the selected input variables. Simple linear models, multiple regression models, feed-forward multilayer perceptron networks etc. were compared in modelling NO_2 concentrations. Pollutant concentrations rarely follow a Normal distribution. NO_2 is no different from the other pollutants, but it can also be modelled using the statistical distributions from the flexible exponential family distribution and it also shares the statistical characteristics found in other pollutants. The exponential family distributions give the much needed flexibility in the construction of such models (Nelder and Wedderburn, 1972).

The GLM model is used to model NO_2 emissions at Eskoms coal fueled power plants in this study.

Methodology

The linear regression and the GLM models are discussed in this section.

Linear regression

This section focuses on models to be used in regression under the Normality assumption of the response variable NO_2. This assumption implies that the emission data is symmetric.

The following model will be fitted on the NO_2 emission data initially. Analysis of Covariance (ANCOVA) is applicable, since the explanatory variables are both continuous and categorical and the response variable is continuous.

$$Y_{pqt} = \beta_0 + \beta_1 x_{pqt} + \beta_2 Age_t + \gamma_p + \alpha_q + \tau_s + \varepsilon_{pqt} \quad (1)$$

where,

Y_{pqt} is the response variable (NO_2 emitted in tons by plant p with abatement filter q and at time t (in years))

β_0 is the intercept

β_1 is the coefficient of the electricity sent out in Gigawatt-hours

β_2 is the coefficient of the age of the power station in years

Age_t of the power plant in years at time t.

x_{pqt} is the amount of electricity sent out in Gigawatt-hours by plant p with filter q at age t

γ_p is the p_{th} plant effect

$_q$ is the q_{th} filter effect

$_s$ is the s_{th} month effect

$_{pqt}$ N(0,^{2})

The model includes all the variables recorded in the study. The group-lasso interaction network variable selection is then selected to try and find a competing model with more variables including interaction terms in the variables mentioned above.

Model selection

Various model variable selection methods exist, such as, among others, subset selection (namely, Best-subset selection, Forward- and Backward-stepwise selection), shrinkage (namely, Ridge Regression, Lasso and Least angle regression) and methods using derived input directions (namely, Principal Components Regression and Partial Least Squares). The group-lasso interaction network (glinternet) is used in this paper to select significant variables (Lim and Hastie, 2015). Lasso regression is selected and used in this paper since it performs both variable selection and regularisation (shrinkage reducing model variance) to enhance predictor accuracy. The method used is an extension of the lasso (least absolute shrinkage and selection operator) variable selection technique (Tibshirani, 1996) and uses a version of the group-lasso to select pairwise interactions and enforce hierarchy (Yuan and Lin, 2006; Bien, Taylor and Tibshirani, 2013). It automatically selects and adds pairwise interactions into the Lasso model. The model selection procedure implies that not all variables may be used in the final model.

A model with less variables- no interaction terms

The glinternet variable selection approach, without interaction terms (only the main effects), is used to select a few significant variables from equation (1) above.

The residuals plot is also used to determine the best fitting model. A constant (homogeneous) residual pattern (constant variance) plot over the predicted values, suggest a good fit or an improvement in the model of interest.

To check if the assumption of normality of data and residuals holds, the box plot, histogram, Kolmogorov-Smirnov test and the quantile-quantile (QQ) plot are used in this study. A symmetric bell-shaped histogram would suggest the data is Normally distributed. The best model is found in the case of the Normal and Lognormal distributions and all with identity link functions as discussed later.

A Model with more terms- including interaction terms

In this section a more complex model is presented using glinternet and allowing for the interaction of variables.

For this model, the selection process will consider all the explanatory variables and all pairwise interaction terms: where the star (*) implies an interaction term. The full model is given as:

$$\eta_{tpqs} = \beta_0 + \beta_1 x_{tpqs} + \beta_2 Age_t + \gamma_p + \alpha_q + \tau_s + \gamma_p * \tau_s + \alpha_q * \tau_s + \gamma_p * \alpha_q + x_{tpqs} * Age_t + x_{tpqs} * \gamma_p + x_{tpqs} * \alpha_q + x_{tpqs} * \tau_s + Age_t * \alpha_q + Age_t * \tau_s + Age_t * \gamma_p,$$

where, for instance,

$x_{tpqs} * Age_t$ is the joint effect of electricity sent out in Gigawatt-hours (by filter q in plant p at given age at t in month s) and Age_t of the power plant at time t (in years).

Other interaction parameters can be interpreted similarly.

Generalised linear models

In a classical regression model with data being Normally distributed, the variance Var(Y)=^{2} is assumed constant. However, in practise, it is common to find data in the form of continuous measurements where the variance increases with the mean (McCullagh and Nelder, 1989). The Lognormal model is one such model.

The Lognormal distribution

If a random variable x is such that $x \sim N(\mu, \sigma^2)$, then under the transformation $y = e^x$ then $Y \sim Lognormal(\mu, \sigma^2) \Leftrightarrow \ln(Y) \sim N(\mu, \sigma^2)$. To fit a Lognormal distribution to a data set, one can firstly log transform the data and then fit a normal distribution to it.

If a random variable Y, with pdf $f(y;\mu,\sigma^2) = \frac{1}{y(2\pi\sigma^2)^{1/2}} \exp\left[-\frac{1}{2\sigma^2}(\ln(y)-\mu)^2\right], -\infty \le \mu \le \infty, y > 0, \sigma > 0$, has a variance that increases with the mean, that is for small , the appropriate variance $Var(Y) = \theta^2[E(Y)]^2 = \mu^2\theta^2$ stabilising transformation would be the logarithm.

The Lognormal model has a variance which increases with the mean. The variance increases with the mean in such a way that the coefficient of variation is a constant. The Lognormal modelling can be used to compensate for such increases with the mean. Also, for small , the log-transformed variable ln(Y) has approximate mean and variance given by

$$E(\ln(Y)) = \frac{1}{n}\sum_{i=1}^{n} \ln(y_i) \approx \ln(\mu) - \frac{\sigma^2}{2}$$

The log transformed variable has variance given as

$$Var[\ln(Y)] \approx \left[\frac{\partial lny}{\partial y}|_{y=\mu}\right]^2 . Var(Y) = \frac{1}{\mu^2} \times \mu^2\theta^2 = \theta^2$$

Since the variance of the Lognormal distribution can be written as $Var(Y) = \mu^2\theta^2$,

where, $E(Y) = u = \exp(\mu + \frac{\sigma^2}{2})$ has $\theta = \sqrt{(\exp(\sigma^2) - 1)}$

The logarithm of the data has a constant variance. Log transforming the data should result in homoscedasticity. Therefore, the GLM Lognormal distribution model will be used to compensate for increases in variance of the emission with increases in mean emissions when such an effect is present in the data.

The exponential family and canonical form

Consider a random variable Y with distribution in the exponential family and pdf f(y;μ) in the standard form:

$$f(y;\mu) = \exp[a(y)b(\mu) + c(\mu) + d(y)]$$

When a(Y)=y, the distribution is said to be in canonical form.

For a distribution to be a GLM, it must have the three components, namely the error distribution, linear predictor and link function (Dobson and Barnett, 2008). The Lognormal distribution is in the exponential family has:

Error Distribution

The Lognormal distribution has independent response variables $Y_1, Y_2, \dots, Y_n$ with $Y_i \sim Lognormal(\mu, \sigma^2)$ with pdf given as

$f(y;\mu) = \exp\left\{\ln(y).\frac{\mu}{\sigma^2} - \frac{\mu^2}{2\sigma^2} - \left[\frac{\ln(y)^2}{2\sigma^2} + \frac{1}{2}\ln(2\pi\sigma^2y^2)\right]\right\}$,

where

$a(y) = \ln(y)$; $b(\mu) = \frac{\mu}{\sigma^2}$; $c(\mu) = -\frac{\mu^2}{2\sigma^2}$; $d(y) = -\frac{\ln(y)^2}{2\sigma^2} - \frac{1}{2}\ln(2\pi\sigma^2y^2)$

The distribution is not in canonical form since a(y)=ln(y).

Linear Predictor

The linear predictor is chosen as for instance

The parameters β and explanatory variable vector X_i are such that

$$\eta_{tpqs} = \beta_0 + \beta_1 x_{tpqs} + \beta_2 Age_t + \gamma_p + \alpha_q + \tau_s,$$

where, $X_i = [1, x_{i1}, x_{i2} \dots, x_{ip}] = [1, x_{tpqs}, Age_t \dots, December]$

Link function

A flexible family of transformations, the power transformations, was introduced by (Box and Cox, 1964). For a given parameter λ, the transformation is defined by

$$g(y) = \begin{cases} \frac{y^\lambda - 1}{\lambda} & for\ \lambda \neq 0 \\ \log(y) & for\ \lambda = 0. \end{cases}$$

The Box-Cox approach is used to estimate the value of λ that will help determine the best link function. According to Myers et al. (2010) the natural values for λ are as follows:

When λ=0 then Log link function
When λ=1 then Identity link function
When λ=1/2 then Square root link function
When λ=-1 then Inverse link function

For the NO_2 data, there exist a monotone link function g such that $g(\mu_i) = \eta_i = x'_i\beta, i = 1, \dots, n$. The choice of a link function can be based on the nature of the data available for the study. The response variable being continuous and positive, the link function is chosen from these

Identity: $g(\mu_{pqt}) = \mu_{pqt} = \beta_0 + \beta_1 x_{pqt} + \beta_2 Age_t + \gamma_p + \alpha_q + \tau_s$ (identity link function, λ=1)

$log: g(\mu_{pqt}) = \log(\mu_{pqt}) = \beta_0 + \beta_1 x_{pqt} + \beta_2 Age_t + \gamma_p + \alpha_q + \tau_s$ (log link function)

Linear regression is a GLM with an identity link

Model selection

Similar,ly to the linear regression, the group-lasso interaction network will be considered in determining models without and with interaction terms, respectively.

Maximum likelihood (ML) is the principal method of estimation used for all GLMs (McCullagh and Nelder, 1989).

In a ML approach, a standard assessment is to compare the fitted model with a fully or saturated specified model (Hardin and Hilbe, 2007). Let β_{max} be the parameter vector of the saturated model and b_{max} be the ML estimator of the β_{max}. The likelihood function of the saturated model evaluated at bmax is $L(b_{max};y)$. For the maximum value L(b;y) of the likelihood function of the model of interest, we have $l(b_{max};y)$ and l(b;y) as the associated log-likelihoods. Such that

$$D = 2\log(\lambda) = 2[l(b_{max};y) - l(b;y)]$$

is the deviance. The deviance for the Lognormal distribution model is given by

$$D = 2\log(\lambda) = \frac{1}{\sigma^2}\sum_{i=1}^{n}(\ln(y_i) - \hat{\mu})^2$$

$$= \frac{1}{\sigma^2}\sum_{i=1}^{n}(\ln(y_i) - \hat{\mu})^2 .$$

A likelihood ratio test (LRT) can be used to perform a hypothesis test on the parameters of interest. To define this test, let M_1 be a GLM with deviance D_1 and p parameters $\beta_1, \ldots, \beta_p$, and let M_2 be a GLM with deviance D_2 and q<p parameters $\beta_1, \ldots\ldots, \beta_q$. Let β be partitioned as $\beta = [\beta^{(1)}, \beta^{(2)}]'$ where, $\beta^{(1)} = \beta_1, \ldots, \beta_q$ and $\beta^{(2)} = \beta_{q+1}, \ldots, \beta_p$. Under the null hypothesis

$H_0: \beta^{(2)} = 0$ $(against\ H_1: \beta^{(2)} \neq 0)$.

Let $l(\hat{\beta};y)$ be the maximum value of the log-likelihood function for M_1 and let $l(\tilde{\beta};y)$ be the value of the log-likelihood function for M_2. The difference of deviances

$$D_2 - D_1 = 2[l(\hat{\beta};y) - l(\tilde{\beta};y)] \sim \chi^2_{p-q},$$

has an approximate χ^2 distribution with p-q degrees of freedom and is known as the Likelihood Ratio Test statistic of the null hypothesis.

Results

The data used in this paper is monthly NO_2 emissions per station, from Eskom, for a maximum period of 108 months (between 2005 and 2014).

Exploratory data analysis

Before any data analysis can be performed, it is important to explore the data in order to know and understand how it is distributed. Graphical display of the data will be done by using the Histogram, Box plot and the QQ plot for the NO_2 emission (in tons). From Figure 1 below, the histogram looks symmetric but is bimodal and hence is not normally distributed (Kolmogorov-Smirnov p-value<0.01). The Box-plot shows that NO_2 emission (in tons) has skewness and kurtosis (skewness value=-0.11 and Kurtosis= -0.94). The Quantile-Quantile plot suggests that NO_2 emission (in tons) is not Normally distributed since data points deviate from a 450 line towards the extremities on each graph.

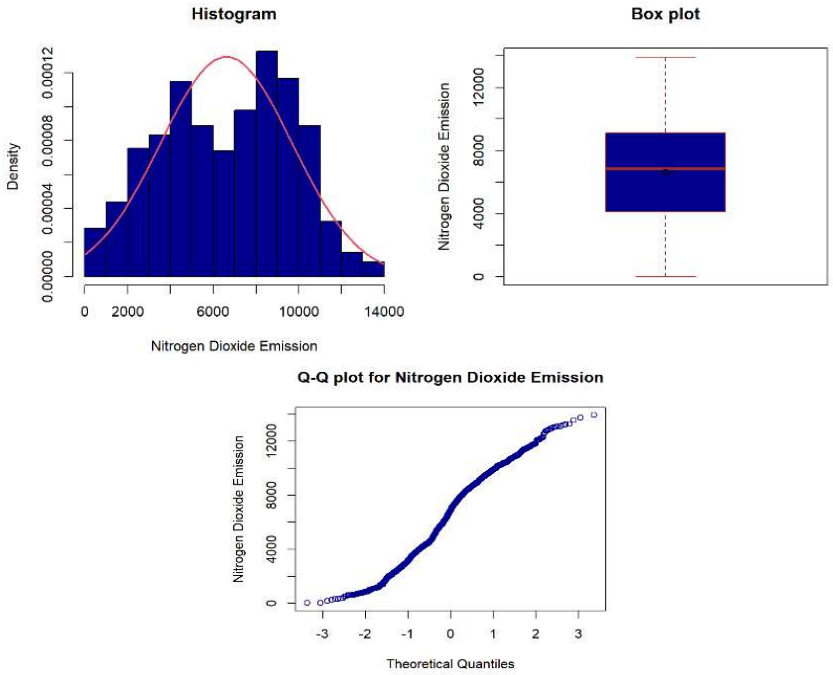

Figure 1: Histogram and Box plot for NO_2 emission (tons)

Efficiency of power stations

Summary statistics on all the power stations used in modelling NO_2 emission (in tons per month) are presented in Figure 2 below

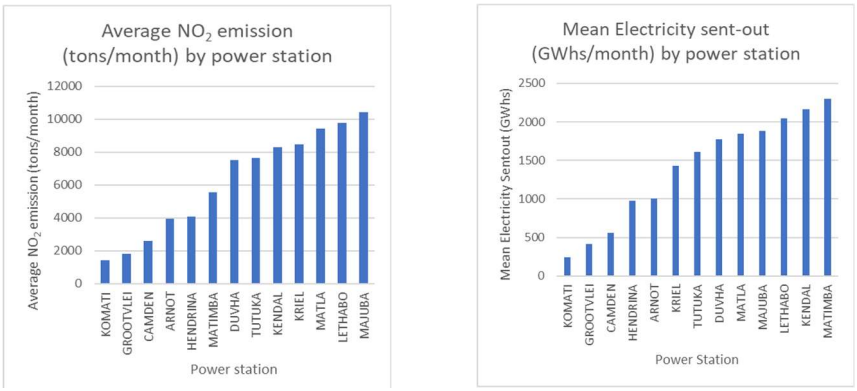

Figure 2: Average NO_2 emission (in tons/month) and average electricity sent-out (GWhs/month), respectively. These graphical representations are given by power station per month

The power station with the lowest average NO_2 emission is Komati with 1422.23 tons per month and the highest is Majuba with 10433.49 tons per month.

Komati power station produced the lowest amount of electricity sent-out (in GWhs) on average per month and Matimba power station produced the highest.

Hendrina is the oldest power station with an age of 44 years and Majuba is the youngest power station with an age of 18 years in year 2014.

Since the efficiency of a power station cannot be measured by observing the amount of NO_2 emission (in tons) alone, the relative nitrogen dioxide (tons/Gigawatt-Hours) was calculated as follows

$$relative\ emission\ (r.e) = \frac{NO_2\ emission\ in\ tons}{electricity\ sent\ out\ in\ Gigawatts\ hours}$$

The power station with the lowest average relative NO_2 emission (tons/GWhs) was taken to be the most efficient of the 13 power stations. Figure 2 shows Matimba with 2.4177 tons/Gigawatt-hours to be the most efficient power station. This suggests that Matimba produces the highest amount of electricity sent out. Kriel is the least efficient power station with 5.96708 (tons/GWhs).

The most efficient month was July with 4.47572 of average relative NO_2 emission (tons/Gigawatt-hours) and January being the least efficient with 4.647 tons/Gigawatt-hours. The month differences are however minimal.

The joint fabric filter, electrostatic precipitators and flue gas condition were associated with the highest efficiency, with 4.27707 tons/GWhs, and electrostatic precipitators are associated with the least efficiency, with an emission of 4.76371 tons/GWhs.

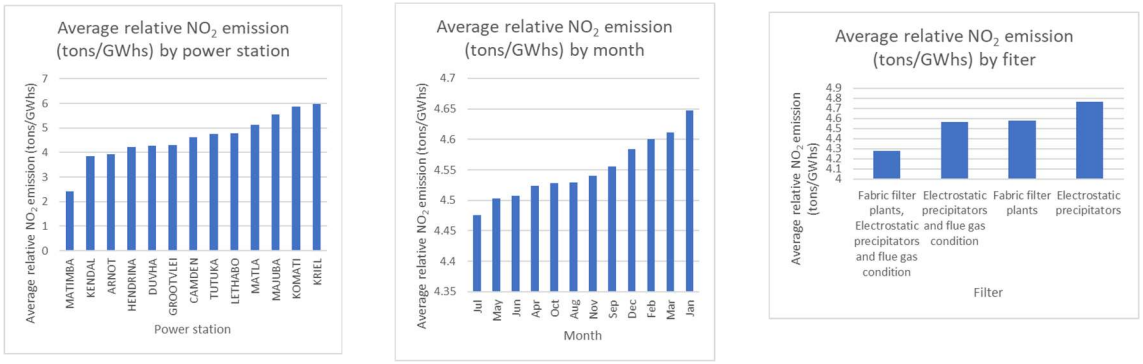

Figure 3: Average Relative NO_2 emission (tons/Gigawatt-hours) by power station, month and filter, respectively

Variable Selection

One of the aims of the paper is to find/select explanatory variables with a significant effect on NO_2 emission at Eskoms power plants.

Test for collinearity (Dependence)

It is important to check for collinearity between some paired continuous explanatory variables before fitting the data to a regression model. The presence of such a relationship would mean that having information about one variable implies that we can predict the other. Thus, both would be trying to explain the same variability for the one response variable.

The variance inflated factors,

$VIF_i = \frac{1}{1-R^2}, i = 1, \ldots, n$ of the explanatory variables will be used to check for collinearity. A value of $VIF_i > 10$ raises concern. R^2 is the coefficient of variation.

As an example, for two variables age (in years) and electricity sent out (in GWhs), we have $VIF_i = 1.71897 < 10$ for each. Which means there is no significant dependence between the two explanatory variables.

The Lasso via hierarchical interactions variable selection

Since no collinearity between variables in the dataset exists, one can start to select a model which includes only the explanatory variables which are significant in determining NO_2 emission (in tons). In determining this, the Lasso (with hierarchical interactions) is used. The information is summarised in Table 1 below.

Table 1: Variables Selected Using Lasso Via Hierarchical Interactions

	Main Effects	Electricity	Age	Filter:A	Filter:B	Filter:C	Filter:D
Electricity	3,950	-	-0,004	0	0	0	0
Age	8,731	-0,004	-	0	0	0	0
Station:ARNOT	-824,742	0,092	-10,366	-41,888	18,235	13,907	9,746
Station:CAMDEN	-1287,188	0,297	9,415	-13,754	8,857	4,529	0,368
Station:DUVHA	-309,013	-0,109	-1,126	5,588	-7,021	2,798	-1,364
Station:GROOTVLEI	-990,525	0,286	-4,843	10,177	-20,790	7,387	3,225
Station:HENDRINA	-289,045	-0,034	-15,204	-27,163	13,327	8,999	4,837
Station:KENDAL	1916,795	-0,828	-62,388	19,048	20,585	-51,729	12,096
Station:KOMATI	-792,174	0,218	-2,250	6,207	7,744	-13,205	-0,746
Station:KRIEL	-657,643	0,331	45,654	-32,388	-30,851	102,580	-39,340
Station:LETHABO	1422,043	0,298	-63,333	-7,894	-6,357	29,097	-14,846
Station:MAJUBA	426,360	1,044	-47,633	71,010	-19,398	-23,725	-27,887
Station:MATIMBA	-1083,048	-1,884	57,082	46,900	48,437	-135,285	39,948
Station:MATLA	-485,036	0,241	28,474	-21,966	-20,429	71,314	-28,918
Station:TUTUKA	-1618,456	0,047	66,520	-13,876	-12,339	-16,667	42,882
Filter:A	-2,051	0	0	-	-	-	-
Filter:B	-3,589	0	0	-	-	-	-
Filter:C	0,739	0	0	-	-	-	-
Filter:D	4,901	0	0	-	-	-	-

Table 1 shows all the coefficients generated by the variable selection process. The table includes the main effects and interaction effects. The first column shows the coefficients of the main effect, and the rest of the columns show the interaction effects. However, not all terms have interaction effects, a 0 indicates such a pair with no interaction effect. The variables amount of electricity sent out (in GWhs), power station used, age of power station (in years), and interaction terms electricity and station, age and station, and station and filter were selected and will be used to produce GLM models for this paper.

In determining the GLMs, a model consisting of only the main effects and without interaction terms will first be considered. The model will be referred to as model I, and is given by

$$Y_{tpq} = \beta_0 + \beta_1 x_{tp} + \beta_2 Age_t + \gamma_p + \alpha_q + \varepsilon_{tpq} \quad (2)$$

The second model with both the main and interaction effects will also be considered and is given by

$$Y_{\boldsymbol{tpq}} = \beta_0 + \beta_1 x_{\boldsymbol{tpq}} + \beta_2 Age_t + \gamma_p + \alpha_q + x_{tpq} \times Age_t + x_{pt} \times \gamma_p + Age_t \times \gamma_p + \gamma_p \times \alpha_q + \varepsilon_{tpq} \quad (3)$$

Generalised Linear Models

Since the results in Figure 1 suggest that NO2 emission (in tons) is not normally distributed, and it is common to find data in the form of continuous measurements where the variance increases with the mean, the Lognormal GLM under model I (model without interaction terms) will be fitted.

Similarly, to the model in equation 2 above, the final model is given by the linear predictor:

$$\eta_{tp} = \beta_0 + \beta_1 x_{pt} + \beta_2 Age_t + \gamma_p,$$

however, the explanatory variable, installed filter, will not be included in this model since it produced parameters with zero values. The model can thus be given as

$$\hat{Y}_{tp} = 6.138 + 0.0008x_{pt} + 0.026Age_t + \hat{\gamma}_p.$$

The plot of residuals versus predicted values, and also, observed values versus predicted values are given in Figures 4 and 5 below, respectively, for the distribution model. Also included in the figures, are the plots for the Normal distribution model with identity link function model. The plots help in assessing the goodness of fit of the models.

The first plots are on residuals versus predicted values (see Figure 4 below)

Below are the plots of the observed versus predicted values (see Figure 5)

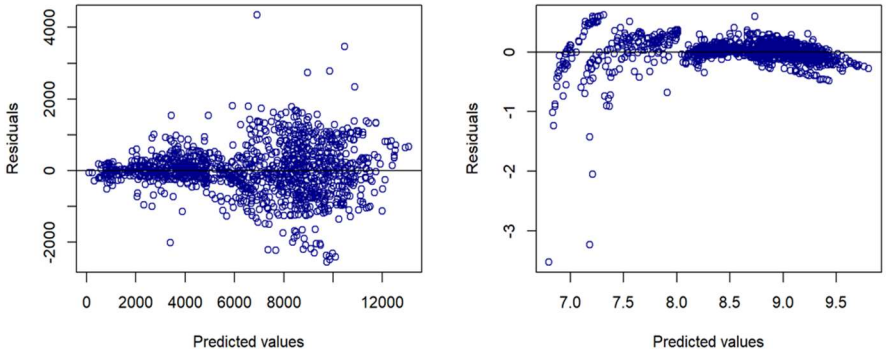

Figure 4: Model I (Model with no interaction terms) residual plots for the Normal and Lognormal models, respectively.

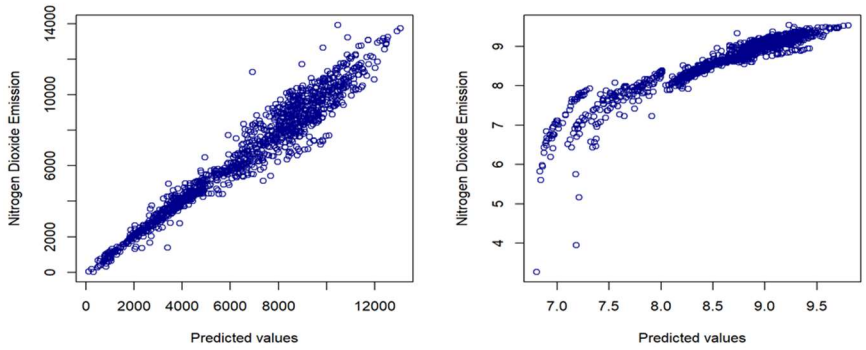

Figure 5: Model I (Model with no interaction terms) actual vs predicted values plots for the Normal and Lognormal models, respectively.

A plot of observed against predicted values again shows the Normal distribution models seems to show an increasing variance with predicted values and hence the model is not very good. On the other hand, the Lognormal model seems to tame the variance behaviour and hence gives the better fit.

A Model with more terms- including interaction terms

In the current section, a model with interaction terms is considered. The resultant model is called Model II and corresponds to the model in equation 2 above.

The Normal model

The final model includes the interaction effects between Electricity and Age, Electricity and Station and Age and station, and explanatory variables electricity sent out (in GWhs), age of power station (in years) and power station used.

The Lognormal model

Similarly, to the normal model above, the final model includes the interaction effects between Electricity and Age, Electricity and Station and Age and station, and explanatory variables electricity sent out (in GWhs), age of power station (in years) and power station used.

For the two models above, the age of the power station is included because of the inclusion of the upper order interaction term Age*station. Also, the interaction term between station and filter, and explanatory variable filter are not included in the final model since they produced coefficients with values of zero.

Thus Model II for the two distributions is given as:

$$Y_{\boldsymbol{tp}} = \beta_0 + \beta_1 x_{\boldsymbol{tp}} + \beta_2 Age_t + \gamma_p + x_{tp} \times Age_t + x_{pt} \times \gamma_p + Age_t \times \gamma_p + \varepsilon_{tp}$$

Below (in Figure 6) are the plots of residuals against predicted values for the Normal and Lognormal distributions under Model II.

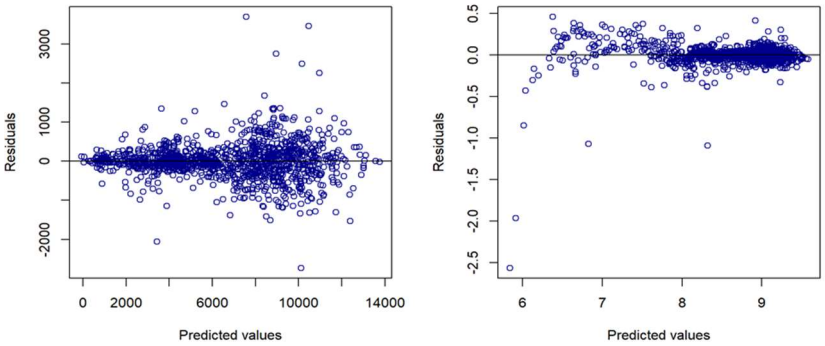

Figure 6: Model II (Model with interaction terms) residual plots for the Normal and Lognormal models respectively.

When the residuals are plotted against predicted values, the Normal model shows an increasing variance with predicted values and hence the model with these interaction terms is also not good. The Lognormal model seems to tame the variance behaviour and hence gives the better fit.

The results of the actual against predicted plots in Figure 7 below also confirm this observation.

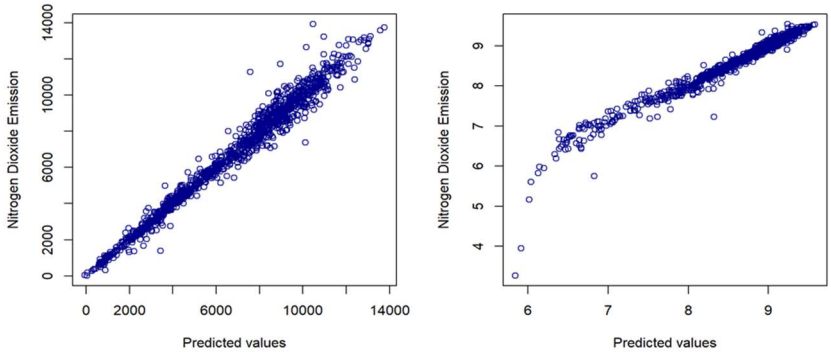

Figure 7: Model II (Model with interaction terms) actual vs predicted values plots for the Normal and Lognormal models, respectively.

Link functions and the deviance

In order to check for a good fit, the deviance was compared to the degrees of freedom. Below are the tables showing the model used, the deviance, degrees of freedom and the associated link functions for the Normal and Lognormal distributions models.

Table 2: Deviances and the different link function for Normal distribution

Model	Degrees of freedom	Link Functions for the Normal models		
		Identity	Log	Inverse
Model I $\eta_{tp} = \beta_0 + \beta_1 x_{pt} + \beta_2 Age_t + \gamma_p$	1281	591964285.93**	674526234.77	Model did not converge
Model II $\eta_{tp} = \beta_0 + \beta_1 x_{tp} + \beta_2 Age_t + \gamma_p + x_{tp} Age_t + x_{tp}\gamma_p + Age_t\gamma_p + \varepsilon_{tp}$	1256	268150806.56**	284590908.65	Model did not converge

Normal distribution

The degrees of freedom for models I and II above are very small compared to their corresponding deviances, that is

$$D_{1i} > DF_1 = 1281 \quad \text{and } D_{2i} > DF_2 = 1256$$

$D1_i$ and $D2_i$ are the deviances for model I and model II, respectively (with i=1 and 2 representing the identity and log link functions, respectively).

DF_1 and $DF2$ are the degrees of freedom for model I and model II, respectively.

This observation suggests that the Normal distribution is not a good fit in modelling NO2 emissions from Eskoms coal fuelled power stations. This observation was checked and confirmed by the use of residual plots and actual versus predicted plots. The identity link function gave the lowest deviance and was hence used.

Lognormal model

Table 3: Deviance and link function for Lognormal distribution

Model	Degrees of freedom	Link Functions for the Lognormal model		
		Identity	Inverse	Lo
Model I $\eta_{tp} = \beta_0 + \beta_1 x_{pt} + \beta_2 Age_t + \gamma_p$	1281	70.42**	80.31	75.7
Model II $\eta_{tp} = \beta_0 + \beta_1 x_{tp} + \beta_2 Age_t + \gamma_p + x_{tp} Age_t + x_{tp}\gamma_p + Age_t\gamma_p + \varepsilon_{tp}$	1256	23.98**	28.05	26.0

All the models from the Lognormal distribution show a good fit to the data since the deviance for each link function is smaller than the degrees of freedom, that is

$D_{1i} < DF_1 = 1281$ and $D_{2i} < DF_2 = 1256$

where,

D_{1i} and D_{2i} are the deviances for model I and model II, respectively (with i=1 and 2 representing the identity and log link functions, respectively).

DF_1 and DF_2 are the degrees of freedom for model I and model II, respectively.

Under the Lognormal model (for both model I and model II), the best fit is with the identity link function since it has the smallest deviance value of the three link functions.

Parameter estimation

Parameters were estimated using ML estimation with Matimba as the basis for comparison since it produced the lowest volumes of average relative NO_2 emissions and hence was the most efficient.

The Lognormal distribution with identity link function (model I detailed results)

Model I: Model with explanatory variables electricity sent out (in GWhs), age of power station (in years) and power station used. Table 4 gives the parameter estimates of the best fitting model I using the Lognormal model as discussed above.

Table 4: Lognormal model *I* (with no interaction terms): Analysis of ML parameter estimates

Parameter		DF	Estimate	Standard Error	Likelihood Ratio 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept		1	6.1380	0.0964	5.9488	6.3271	4051.04	<.0001
Electricity_Sentout		1	0.0008	0.0000	0.0008	0.0009	824.01	<.0001
Age		1	0.0260	0.0026	0.0209	0.0312	98.42	<.0001
Power_Station_Effect	ARNOT	1	0.3132	0.0610	0.1935	0.4329	26.34	<.0001
Power_Station_Effect	CAMDEN	1	0.2735	0.0640	0.1480	0.3990	18.27	<.0001
Power_Station_Effect	DUVHA	1	0.5529	0.0387	0.4769	0.6288	203.89	<.0001
Power_Station_Effect	GROOTVLEI	1	0.1369	0.0670	0.0054	0.2684	4.17	0.0412
Power_Station_Effect	HENDRINA	1	0.3203	0.0647	0.1934	0.4473	24.48	<.0001
Power_Station_Effect	KENDAL	1	0.5276	0.0321	0.4646	0.5906	269.93	<.0001
Power_Station_Effect	KOMATI	1	-0.1939	0.0745	-0.3400	-0.0479	6.78	0.0092
Power_Station_Effect	KRIEL	1	0.8213	0.0491	0.7250	0.9176	279.81	<.0001
Power_Station_Effect	LETHABO	1	0.7189	0.0328	0.6545	0.7833	479.45	<.0001
Power_Station_Effect	MAJUBA	1	1.1764	0.0409	1.0963	1.2565	829.25	<.0001
Power_Station_Effect	MATLA	1	0.6919	0.0392	0.6150	0.7689	311.02	<.0001
Power_Station_Effect	TUTUKA	1	0.8252	0.0374	0.7518	0.8986	486.09	<.0001
Power_Station_Effect	MATIMBA	0	0.0000	0.0000	0.0000	0.0000	.	.
Scale		1	0.2331	0.0046	0.2244	0.2424		

Table 4 shows the ML parameter estimate of electricity sent out (in GWhs) of 0.0008. This means that an increase in electricity sent out by 1 Gigawatt-hour will increase the log NO_2 emission in log tons by 0.0008 (equivalent to 1.0008 tons). Other log tons estimates will be similarly interpreted.

An estimate with a positive value for the plant coefficient means the associated power station variable in the model has the effect to produce log NO_2 emission exceeding those of the basis, Matimba, by the estimated value. A negative value means the basis (Matimba) effect exceeded the log NO_2 emission of

the associated power station by the value of the estimate. The lowest plant coefficient implies the lowest impact on emission (in log tons of NO2) having taken account the other variables in the model. The highest plant coefficient implies the highest log NO_2 emission impact.

Komati, Grootvlei and Camden produce less electricity and hence are expected to produce less NO_2 emissions.

According to the power plant parameter estimates in Table 4, Komati (with log emission level of 0.1939 log tons less than Matimba) has the least impact of the 13 power stations. It has the lowest parameter estimate (and the only estimate with a negative value). Majuba (with 1.1764 log tons more than Matimba) has the greatest impact in increasing emissions. The parameters are interpreted in the presence of other variables in Model I.

The Lognormal distribution with Identity link function: Model II (model with interaction terms)

The parameter estimates for the best Model II are given in table 5. This model consists of the explanatory variables electricity sent out (in GWhs), age of power station (in years) and power station used, and the interaction terms electricity*age, electricity*station and age*stationIn Table 5 above, the ML coefficient of electricity sent out (in GWhs) is 0.0007. This means that an increase in electricity sent out by 1 Gigawatt-hour will increase the log NO_2 emission in log tons by 0.0007 units (equivalent to 1.0007 tons). On the other hand, an increase of age by a year will increase log NO_2 emission by 0.0298 log tons (equivalent to 1.0302 tons).

Table 5 gives the power station effect in the presence of other variables in the Lognormal model. According to the Lognormal Model II, the power stations Arnot, Hendrina, Camden, Grootvlei, Tutuka, Komati and Kriel had less effect on emissions compared to Matimba since these had negative parameter estimates. This is happening when interaction effects are allowed for. Arnot (with 0.9759 log tons less than Matimba) had the least effect from the 13 power stations followed by Hendrina (with 0.7919 log tons less than Matimba). Duvha, Matla, Majuba, Lethabo and Kendal had the greatest effect in increasing emissions compared to Matimba, with Kendal (emission level of 1.6753 log tons more than Matimba) contributing the greatest effect on emissions of all the 13 power stations.

Since the interaction of electricity sent out (in GWhs) and age produced a very small value of the estimate such that the software package used cannot display it but its sign only, we can only conclude that the joint increase in electricity sent out by 1 Gigawatt-hour and increase in age by a year will decrease the log NO_2 emission in log tons by a value less than 0.0001 units.

Taking a closer look on the interaction term: electricity*station

For the interaction term electricity *station, the least effect from the 13 power stations comes from the interaction term electricity*Kendal (with only 0.0001 log tons less than electricity*Matimba) and the interaction of the electricity variable with Komati power station has the greatest effect to increase emissions significantly (with 0.0051 more log tons when compared to electricity*Matimba).

Komati, Grootvlei and Camden produce less electricity and hence are expected to produce less NO_2 emissions. However, the emissions are disproportionately higher.

For the effect age*station, Komati, Kendal, Camden, Grootvlei, Lethabo, Hendrina, Duvha and Majuba have interaction with age coefficients to reduce emission impact since they all have negative interaction coefficients when compared with the basis, age*Matimba. Age interaction with, Kriel, Arnot, Matla and Tutuka contribute to increasing emissions since the coefficients are all positive. The interaction term age*station has Komati (with 0.0575 log tons less than age*Matimba) leading to the least impact on emission. Age interaction with Tutuka leads to the greatest emission impact (with 0.0404 more log tons compared to age*Matimba). Generally, the older plants give more emissions. Tutuka produces more emissions than expected given its age.

Criteria for assessing goodness of fit: Selecting the best model.

One can now determine if the addition of interaction terms produced a better fit or not when compared to the model with less terms (no interaction effects).

Lognormal model with identity link function

Let D_1 and D2 be the deviances for models I and II, respectively, such that

$D_1 = 70.4203$ with $degrees\ of\ freedom = 1281$

and

$D_2 = 23.9824$ with $degrees\ of\ freedom = 1256$

Now, under the null hypothesis given as

$$H_0: \underline{\beta}^{(2)} 0\ against\ H_1: \underline{\beta}^{(2)} \neq 0$$

We have

$D_1 - D_2 = 70.4203 - 23.9824 = 46.4379 > 37.65 = \chi^2_{1281-1256} = \chi^2_{25}$

This suggest that the null hypothesis $H_0: \underline{\beta}^{(2)} = 0$ will be rejected at α=0.05 and we can conclude that the addition of the interaction terms is significant in predicting the emission of NO_2 and thus model II can be used in predicting NO_2 emission and can be regarded as the best fit of the two.

Evaluating the predictive models (RMSE, MAPE and MAE)

In addition to the residuals plots above, prediction evaluation metrices are presented to confirm the fitting model. Table 6 below shows the root mean squared error (RMSE), mean absolute percentage error (MAPE) and the mean absolute error (MAE) for the two models, Normal and Lognormal distributions.

From Table 6 above, the MAPE for the Lognormal model (with a value of 0.86%) is lower than that of the Normal distribution (with a value 5.34%). This suggests the Lognormal model is a better fit compared to the Normal model. This is supported by the results of the RMSE and MAE i.e. for the Lognormal model II, MAE has a lower value of 0.0653 log tons (equivalent to 1.0675 tons) compared to 296.1763 tons of the Normal distribution model II.

Table 6: Prediction evaluation metrices for the Normal and Lognormal models with interaction terms (Model II)

Distribution model II	RMSE	MAPE	MA
Normal	454.8697	0.0534	296.1
Lognormal	0.1360	0.0086	0.06

Discussion

In a classical regression model, the variance is assumed a constant and the data is assumed to be normally distributed. However, in practice, it is common to find data in continuous measurements where the variance increases with the mean (McCullagh and Nelder, 1989). In such cases, a Lognormal GLM could be used. Diagnostic plots suggest an increasing variance with an increasing mean for this data set. The data set obeys the constant coefficient of variation assumption.

The results of the linear regression model suggest that NO2 emission data is not Normally distributed. This is supported by the results from the histogram, box plot. The Lognormal distribution models are also fitted to the data. The best link function is the identity link as evidenced by the smallest deviance compared to the log and inverse link functions. Intermediate results in comparisons of the Lognormal model with identity link function and linear regression model, using the residuals plots and actual versus predicted plot, indicate that, the Lognormal model is better as it produced plots that showed improved variance behaviour that is now constant. It can be concluded that, the GLM model is a better model than the linear regression model in explaining and predicting NO_2 emission data from Eskoms coal-fuelled power stations.

The identification of significant variables contributing to high emissions is essential in the monitoring and managing of emissions. The interaction terms electricity*station, Age*station and variables electricity sent out (in GWhs), age of power station (in years), power station used can be used in describing and predicting NO_2 emissions from Eskoms coal fuelled power stations.

To enhance research on NO_2 emissions from Eskom coal fuelled power stations, it would be beneficial to add the amount and quality of coal used in the generation of electricity as some of the explanatory variables. For future studies, the researchers would like to compare two GLM distributions models that obey the constant coefficient of variation assumptions namely, Lognormal and Gamma models.

Conclusion

This paper discusses the use of GLMs in the modelling of emission data from the 13 Eskoms coal-fuelled power stations. GLM distribution models, namely the Normal and the Lognormal, were constructed and compared. Each distribution model was divided into two, one without (Model I) and the other with interaction terms (Model II), respectively, by making use of group-lasso interaction network (glinternet) variable selection method. This was done to determine if addition of interaction effects in the models is significant or not. The deviance was then used to determine the best link function between the identity, log and inverse for Model I and Model II. The identity link function was deemed the most appropriate for the given dataset. In the case of the Normal GLM models, the deviance had values that were very large compared to their corresponding degrees of freedom, suggesting that the Normal distribution models (and thus the linear regression models) are not a good fit for the data. This is expected as it is common to have continuous data, including emission data, that does not obey the Normality assumption (McCullagh and Nelder, 1989). We can, therefore, conclude that the linear regression model is not a good fit for the NO_2 emission data. For the Lognormal distribution model, the addition of interaction terms was significant. The main contribution of this paper is to demonstrate the GLMs flexibility offered by the link functions to transform the data compared to the limited classical linear regression when modelling NO_2 emission data (Nelder and Wedderburn, 1972). The modelling helps in coming up with better models to explain Eskom emmissiom data, such the NO_2 emission data. The study is useful to power utilities such as Eskom in the monitoring and management of emissions to meet the regulations and thus manage the emission to minimise the exposure of high NO_2 emissions to humans and the environment.

Table 5: Lognormal model II: Analysis of ML Parameter Estimates

Analysis Of Maximum Likelihood Parameter Estimates								
Parameter		DF	Estimate	Standard Error	Likelihood Ratio 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept		1	6.9384	0.4199	6.1148	7.7620	273.05	<.0001
electricity		1	0.0007	0.0002	0.0004	0.0010	17.57	<.0001
Age		1	0.0298	0.0167	-0.0031	0.0626	3.16	0.0754
station	ARNOT	1	-0.9759	0.3318	-1.6267	-0.3251	8.65	0.0033
station	CAMDEN	1	-0.6835	0.4174	-1.5021	0.1351	2.68	0.1015
station	DUVHA	1	0.0676	0.3965	-0.7101	0.8453	0.03	0.8646
station	GROOTVLEI	1	-0.6122	0.6821	-1.9502	0.7258	0.81	0.3695
station	HENDRINA	1	-0.7919	0.4624	-1.6989	0.1150	2.93	0.0868
station	KENDAL	1	1.6753	0.2728	1.1403	2.2104	37.72	<.0001
station	KOMATI	1	-0.2227	1.0517	-2.2854	1.8401	0.04	0.8323
station	KRIEL	1	-0.0715	0.3220	-0.7029	0.5600	0.05	0.8244
station	LETHABO	1	1.3508	0.2938	0.7746	1.9270	21.14	<.0001
station	MAJUBA	1	0.8762	0.3054	0.2772	1.4752	8.23	0.0041
station	MATLA	1	0.1379	0.3545	-0.5575	0.8332	0.15	0.6973
station	TUTUKA	1	-0.5523	0.2839	-1.1091	0.0046	3.78	0.0517
station	MATIMBA	0	0.0000	0.0000	0.0000	0.0000	.	.
electricity*Age		1	-0.0000	0.0000	-0.0000	0.0000	3.08	0.0791
electricity*station	ARNOT	1	0.0007	0.0002	0.0004	0.0010	21.05	<.0001
electricity*station	CAMDEN	1	0.0032	0.0002	0.0028	0.0035	384.35	<.0001
electricity*station	DUVHA	1	0.0002	0.0001	0.0000	0.0004	4.74	0.0295
electricity*station	GROOTVLEI	1	0.0024	0.0002	0.0020	0.0029	128.86	<.0001
electricity*station	HENDRINA	1	0.0009	0.0002	0.0005	0.0012	19.75	<.0001
electricity*station	KENDAL	1	-0.0001	0.0001	-0.0003	0.0001	0.93	0.3355
electricity*station	KOMATI	1	0.0051	0.0003	0.0044	0.0057	209.71	<.0001
electricity*station	KRIEL	1	0.0004	0.0001	0.0002	0.0006	12.40	0.0004
electricity*station	LETHABO	1	0.0000	0.0001	-0.0002	0.0002	0.00	0.9677
electricity*station	MAJUBA	1	-0.0000	0.0001	-0.0002	0.0002	0.00	0.9792

References

Anand, S., Varma, K. and Srimurali, M. (2013) Concentration of Nitrogen Dioxide Estimation from Modeled NOX of a Thermal Power Plant, Journal of Environmental Science, Toxicology and Food Technology, 6(3), pp. 0811.

Bien, J., Taylor, J. and Tibshirani, R. (2013) A lasso for hierarchical interactions, The Annals of Statistics, 41(3). doi:10.1214/13-AOS1096.

Box, G.E.P. and Cox, D.R. (1964) An Analysis of Transformations, Journal of the Royal Statistical Society: Series B (Methodological), 26(2), pp. 211243. doi:10.1111/j.2517-6161.1964.tb00553.x.

Capilla, C. (2014) Multilayer perceptron and regression modelling to forecast hourly nitrogen dioxide concentrations, WIT Transactions on Ecology and The Environment, 183, pp. 3948.

Dobson, A.J. and Barnett, A.G. (2008) An introduction to generalized linear models. 3rd edn. Edited by B.P. Carlin et al. CHAPMAN & HALL/CRC: Texts in Statistical Science Series.

Eskom (2011) COP17 fact sheet: Air quality and climate change. Available at: http://www.eskom.co.za (Accessed: 10 December 2015).

Eskom (2016) 2014s best performing return to service project globally - Camden. Available at: http://www.eskom.co.za/news/Pages/Feb11.aspx (Accessed: 21 December 2017).

Georgopoulos, P.G. and Seinfeld, J.H. (1982) Statistical distributions of air pollutant concentrations, Environmental Science & Technology, 16(7), pp. 401A-416A. doi:10.1021/es00101a727.

Hardin, J.. and Hilbe, J.. (2007) Generalized linear models and extensions. 2nd edn. StrataCorp LP.

Kukkonen, J. (2003) Extensive evaluation of neural network models for the prediction of NO2 and PM10 concentrations, compared with a deterministic modelling system and measurements in central Helsinki, Atmospheric Environment, 37(32), pp. 45394550. doi:10.1016/S1352-2310(03)00583-1.

Lim, M. and Hastie, T. (2015) Learning Interactions via Hierarchical Group-Lasso Regularization, Journal of Computational and Graphical Statistics, 24(3), pp. 627654. doi:10.1080/10618600.2014.938812.

McCullagh, P. and Nelder, J.A. (1989) Generalized linear models. 2nd edn. London: Chapman and Hall.

Myers, R.H. et al. (2010) Generalized linear models: with applications in engineering and the sciences. John Wiley & Sons.

Nelder, J.A. and Wedderburn, R.W.M. (1972) Generalized Linear Models, Journal of the Royal Statistical Society. Series A (General), 135(3), p. 370. doi:10.2307/2344614.

Perez, P. and Trier, A. (2001a) Prediction of NO and NO2 concentrations near a street with heavy traffic in Santiago, Chile, Atmospheric Environment, 35(10), pp. 17831789. doi:10.1016/S1352-2310(00)00288-0.

Perez, P. and Trier, A. (2001b) Prediction of NO and NO2 concentrations near a street with heavy traffic in Santiago, Chile, Atmospheric Environment, 35(10), pp. 17831789. doi:10.1016/S1352-2310(00)00288-0.

Pretorius, I. et al. (2015) A perspective on South African coal fired power station emissions, Journal of Energy in Southern Africa, 26(3), pp. 2740.

Singh, K.P. et al. (2001) Mathematical modeling of environmental data, Mathematical and Computer Modelling, 33(6 7), pp. 793800. doi:10.1016/S0895-7177(00)00281-8.

Tibshirani, R. (1996) Regression Shrinkage and Selection Via the Lasso, Journal of the Royal Statistical Society: Series B (Methodological), 58(1), pp. 267288. doi:10.1111/j.2517-6161.1996.tb02080.x.

Wellenius, G.A., Schwartz, J. and Mittleman, M.A. (2015) Health and the environment: addressing the health impact of air pollution, Draft resolution proposed by the delegations of Albania, Chile, Colombia, France, Germany, Monaco, Norway, Panama, Sweden, Switzerland, Ukraine, United States of America, Uruguay and Zambia. Sixty-Eighth World Health Assembly. Agenda item, 14, p. A68.

World Health Organization (2013) Health Effects of Particulate Matter: Policy implications for countries in eastern Europe, Caucasus and central Asia. Available at: https://apps.who.int/iris/handle/10665/344854.

Yuan, M. and Lin, Y. (2006) Model selection and estimation in regression with grouped variables, Journal of the Royal Statistical Society: Series B (Statistical Methodology), 68(1), pp. 4967. doi:10.1111/j.1467-9868.2005.00532.x