NEW TRENDS IN MODELLING CLIMATE CHANGE IN THE ERA OF BIG DATA

[1] Osuolale Peter Popoola, [2] Nicholas Nsowah Nuamah

[1] Maths and Statistics Department, The Ibarapa Polytechnic, Eruwa Oyo stste Nigeria and
[2] Regent University College of Sience and Technology Accra Ghana

Corresponding Author: Osuolale Peter Popoola, osuolalepeter@yahoo.com

ABSTRACT: Big data is data sets that are so voluminous and complex that traditional data processing application software are inadequate to deal with. It is typically characterized by the so called, seven "V's" namely; volume, velocity, variety, Value, Veracity, variability and validity. Big Data can be thus defined as very high volume, velocity and variety of data that require a new high-performance processing. Thus, this research work survey different technique for handling big data for modelling climate change. Various statistical models were examined and X-ray. The survey shows that ANCOVA (analysis of covariance) is the appropriate technique for handling big data. ANCOVA contains a mixture of qualitative variables associated with analysis of variance (ANOVA) and the quantitative variables associated with regression analysis. ANCOVA is the meeting point under the umbrella of analysis of variance and regression techniques.

KEYWORDS: Big Data, Pooled Data, Analysis of Variance, Analysis of Covariance and Multiple Regression.

1.INTRODUCTION

Big data is data sets that are so voluminous and complex that traditional data processing application software are inadequate to deal with them. It is typically characterized by the so called, seven "V's" namely; volume, velocity, variety, Value, Veracity, variability and validity. Therefore, big data is defined as "the variable information assets characterized by such a High Volume, Velocity and Variety to require specific Technology and Analytical Methods for its transformation into Value whose veracity can be validated [C+16] By this definition, Big data can be described by the following seven characteristics or properties which are volume, variety, velocity, veracity, value, variability and validation. Big data require specific Technology and Analytical Methods for its transformation into Value" has been used [ES07].

Volume: The quantity of generated and stored data is large-scale. No sampling but observes and tracks what happens. The size of the data determines the value and potential insight- and whether it can actually be considered big data or not.

Variety: The type and nature of the data are text, images, audio, video; plus integrating multiple data sources to complete missing data. This helps people who analyze it to effectively use the resulting insight.

Velocity: Data is rapidly changing, like an optimization problem in dynamic environment but often available in real-time. In this context, the speed at which the data is generated and processed to meet the demands and challenges that lie in the path of growth and development is important.

Value: The value of data is the objective of the big data analytics, like the fitness or objective function in an optimization problem [W+14]. The value of information should drive the investment in Big Data and the collection of it, for its own sake.

Variability: Inconsistency of the data set can hamper processes to handle and manage it.

Veracity: Data is inconsistent and/or incomplete, like an optimization problem with noise or approximation [W+14]. The quality of captured data can vary greatly, affecting accurate analysis [Hil15].

Validation: The validity of the data should be ascertained before the information derived thereof can be considered appropriate.

The Vs have been expanded to other complementary characteristics of big data [***18, Gri16].

1. Machine learning: big data often doesn't ask why and simply detects patterns.
2. Digital footprint: big data is often a cost-free byproduct of digital interaction.

Climate change is already underway with consequences that must be faced today as well as tomorrow. Evidence of changes to the Earth's physical, chemical, and biological processes is now evident on every continent of the world. The question that is yet to be answered is how do we model the effect of climate change in the era of big data? What is referred to as Big Data is relative and depends on the capabilities of the users and their tools, and expanding capabilities make big data a moving target. "For some organizations, facing hundreds of gigabytes of data for the first time may trigger a need to reconsider data management options. For others, it may take tens or hundreds of terabytes before data size becomes a significant consideration."[Hil15]. However, the current data processing capacity to handle volume of data are beyond the computing ability of traditional computational models [BC11, W+14]. Climate models use quantitative methods to simulate the interactions of the important drivers of climate, including atmosphere, oceans, land surface and ice. They are used for a variety of purposes from study of the dynamics of the climate system to projections of future climate. It is often convenient to regard climate models as belonging to one of four main categories, which are

- energy balance models (EBMs);
- one dimensional radiative-convective models (RCMs);
- two-dimensional statistical-dynamical models (SDMs);
- three-dimensional general circulation models (GCMs).

With the increase of (big and unstructured) data collected every day in many disciplines, the amount of available data is growing constantly and exponentially. The world's technological per-capita capacity to store information has roughly doubled every 40 months since the 1980s; as of 2012, every day 2.5 exabytes (2.5×10^{18}) of data are generated. By 2025, IDC predicts there will be 163 zettabytes of data (wikipedia.org/wiki/Bigdata). Today there are several methods of building models based on the data types, uses and scientific substance. The basis for their classification may be on the method of pooling the data into one cell. In considering the method to use in pooling the data for modelling, some questions arise and have to be borne in mind First, to what extent does the result of the model of the pooled data, correspond to the true models, that is, the ones which would have been got for each year, Secondly, which method enables us to determine the existence of structural changes and also to consider these changes in the model? Thirdly, which method could be applied to the pooled data so as to obtain an unbiased estimate of the general model? Fourthly, which statistical paradox arises in those methods which don't provide unbiased estimates? Is the paradox the result of stochastic variables or are they characteristics of a given process?

Fifthly, if the statistical paradox are the characteristics of a given process then in which condition do they arise and in which they do not?

Lastly, but not the least, how best can the parameters of the model be estimated so that the model could be used for statistical analysis and forecasting?

To provide answers to all the above questions, this research shall survey different models available and reccommend appriopriate models for modelling climate change in the era of climate change.

2.GENERAL INFORMATION

2.1. Evolution of Statistics

Statistics development was on two parallel paths:

1. development of official statistics,
2. development of statistics as an academic discipline in the universities.

The ***first path*** can be traced back to the conduct of censuses in Biblical times and the activities of government statistical offices. Rulers of ancient Babylonia, Egypt and Rome and most civilised countries from earliest times gathered detailed information of populations and resources for the purposes of levying taxes to maintain the state and the court and also ascertaining manpower and material strength of the nation for military and fiscal reasons.

The word “Statistics” was first used by a German Professor, Gottfried Achenwall, about 1770. It was then defined as “the Science that teaches us what is the political arrangement of all the modern states of the new world”. Dr. E.A.W. Zimmerman introduced the word "Statistics" into England. As it can be seen, the word was originally referred to as collections of facts (not necessarily numerical) about the state, or the people who composed the state. The facts were then given in a verbal form.

The ***second path*** of development of statistics which is the development of statistics as an academic discipline in the universities can be traced from the work of Pascal and Bernoulli in the 17th Century. Up to the last decade of the 1880s, the two paths of academic statistics and official statistics developed in parallel with little contact between them. At an early stage, the teaching of statistics was captured by mathematicians who were academic statisticians and now Statistics, like physics and economics, undeniably makes heavy and essential use of mathematical tools. At this stage of statistical development, the approach in practice is basically, a complex interplay between the data and a mathematical model. Statistics begins with statement of a problem which has risen in fields such as economics, sociology, biology, health, etc. In statistics, the problems of data, uncertainty and scientific inference are central.

2.2. Linear Models – Interplay of Statistics and Mathematics

A statistical model is a formalized expression of theory or the casual situation that is regarded as having generated observed data. There are two types of statistical models, namely, linear model (linear in parameters) and non-linear model (non-linear in parameters). Linear models are the most widely-used in analyzing the result of an investigation.
A linear model is given by the equation

$$Y = X\beta + \varepsilon \qquad (1)$$

where β is the vector of unknown coefficient, β_1, β_2, ..., β_p, called parameters, which control the behavior of the model; ε is a vector of random errors.

Type of Linear Models

When the values of X matrix can take a continuum of values, Model 1.1 is known as functional model. For example, supposed we wish to fix or explain intelligence quotient(Y) of students in a school by means of model on age (X_1), height (X_2) and weight (X_3). Since the X values can take a continuum of values, we have a functional model.

Classificatory Model

When the X matrix consists of factors of classification and therefore takes discrete values, model 1.1 is called classificatory model. That is, the individuals with Y-values which have to be fitted by the model may be classified according to factors of classification such as color, religion and type. Suppose students whose intelligence quotient (Y) we wish to fit are classified according to sex (X_1), occupation of father or guardian (X_2) and ethnicity (X_3); the X values take discrete values.

Model with functional and classificatory portions

A more general class of linear models has both functional and classificatory components. An example is when we wish to fit or explain intelligence quotient (Y) of students in a school by means of model on age (X_1) and sex (X_3). We observe that X_1 can take a continuum of values while X_2 if a factor of classification and therefore takes discrete values.

Methods of linear models

The methods often encountered in practice for linear models are the ordinary regression analysis (ORA) for functional models and the Analysis of Variance (ANOVA) for classificatory models. For models with both classificatory and functional components, the Analysis of Variance (ANOVA) is used.

For example, suppose we have a variable Y to be explained in terms of variables X_1, X_2, ... X_p. Each of the variable Y, X_1, X_2, ..., X_p are observed on n units or entitles or individuals. The resulting data set may be represented as an n x 1 vector Y, $Y^1 = (Y_1, Y_2, ..., Y_n)$ which is to be explained by means of the columns of X matrix of order n x p.
Where:

$$X = \begin{pmatrix} X_{11} & X_{12} & \dots & X_{1p} \\ X_{21} & X_{22} & \dots & X_{2p} \\ . & . & & \\ . & . & & \\ . & . & & \\ X_{n1} & X_{n2} & \dots & X_{np} \end{pmatrix} = $$
$$= (X_1, X_2, ..., X_p)$$

and where, X_1, X_2, ..., X_p are *n* x 1column vectors. The matrix X is called the design matrix or model matrix and vector Y is the vector of observations.
Consider data on climate for *n* stations or grid observed over a period of *T* years. Define Y_{it} and X_{git} as the value of dependent variable and g^{th} independent variable, respectively, of the i^{th} unit (i = 1, 2, ..., n) of t_{th} year (t = 1, 2, ..., T). Then the observations at a given time *t* may be written as:

$$Y_t = X_t\beta + \varepsilon$$

where

$$Y_t = \begin{pmatrix} Y_{1t} \\ Y_{2t} \\ \vdots \\ Y_{nt} \end{pmatrix} \qquad (2)$$

$$X_t = \begin{pmatrix} X_{11\,t} & X_{21\,t} & \dots & X_{kit} \\ X_{12t} & X_{22t} & \dots & X_{k2t} \\ \vdots & \vdots & \vdots & \vdots \\ X_{1nt} & X_{2nt} & \dots & X_{knp} \end{pmatrix} \qquad (3)$$

$$\beta = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{pmatrix} \qquad (4)$$

$$\varepsilon_t = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_{nt} \end{pmatrix} \qquad (5)$$

2.3. Modelling Pooled Cross-section and Time Series Data

Combining cross section and time series data is common in climate change modelling. The various meteorological stations collect data over years.

One of the methods which is applied before the least squares method, namely, the mean approach (averaging the data over units or over time before regressing) has been extensively studied [Kra81, Nua86, Nua92, Nua00]. It was observed that the regression parameter estimates of the mean approach may not be the same as those of the individual years even if they are the same in all the years. The difference may be significant either in magnitude, direction or both and this may give rise to a statistical paradox. The paradox is that the values of the regression coefficients of the pooled data by the mean approach could exceed the range of values of the corresponding regression coefficients obtained by the classical regression equations for the individual years. By reexpressing the normal equations of the mean approach model the source of this paradox was found to be due to the presence of autocovariances and drift covariances inherent in that approach [Kra81].

The classical pooling approach considers unit-time as points and uses ordinary least squares for model fitting. It has been studied in most cases in econometrics text books and journals. Greene has given a theoretical discussion in his classic text book on Econometrics [Gre93]. In that book, emphasis is on time series analysis while studying cross-sectional heteroscedasticity. Considering the situation when emphasis is on cross-sectional analysis while studying time series heteroscedast.

Discussion of asymtotic properties are coded in terms of n→∞ and or T → ∞ [Gre93]. Whichever one applies will make a difference in a particular context. The asymtopic results we obtained in this paper are with respect to n → ∞ . Assume in this paper that the parameters β_{gt} are constant across time t, that each cross sectional unit is observed the same number of times in all the time pariod and that n is fixed, the model for the classical pooling may then be defined as:

$$Y=X\,\beta+\varepsilon \qquad (6)$$

Where:

Y is the vector of the dependent variable of order *N x* 1;

X is the matrix of the independent variables of order *N* x *k*;

β is the vector of the regression coefficients of order *k* x 1;

ε is the random error vector of order *N* x 1.

$(N=nT=N_1+N_2+\ldots+N_T)$

The data in model (1) have been stacked to obtain *h* = 1,2, ..., *N* observation.

The error structure for Model (1) is a simple one because the classical pooling approach considers the whole set of cross section data as if there were one effect that could fit all time series in the pool. In effect, all the effects of time are indeed captured in that error which cannot be decomposed any further. Therefore, the basic assumptions of model (1) in relation to the error term may be stated as ([Nua92]).

$$\Omega_1 = \begin{cases} E(\varepsilon) = 0 & \text{for all } I \qquad (7) \\ \operatorname{cov}(\varepsilon_{it}, \varepsilon_{i't}) = \begin{cases} \sigma^2_{\varepsilon i} & \text{for } i = i' \\ 0 & \text{for } i \neq i' \end{cases} & \\ \operatorname{cov}(\varepsilon_{it}, \varepsilon_{is}) = \begin{cases} \sigma^2_{\varepsilon t} & \text{for } t = s \\ 0 & \text{for } t \neq s \end{cases} & \qquad (8) \\ \operatorname{cov}(\varepsilon_{it}, \varepsilon_{i's}) = 0 & \text{for } i \neq i';\ t \neq s \qquad (9) \\ \operatorname{cov}(\varepsilon_{it}, X_{git}) = 0 & \text{for any } i, t \qquad (10) \end{cases}$$

The parameters of model (6) are ordinarily unknown and must be estimated from the sample data. The ordinary least squares estimate of the regression parameters are defined by:

$$\beta = (X'X)^{-1}\,XY \qquad (11)$$

Krastin concluded that unlike the mean approach, no paradox can arise in classical pooling [Kra81]. The paradox here is that the values of the regression coefficients of the pooled data could exceed the range of values of the corresponding regression coefficients obtained by the classical regression equations for the individual years. This conclusion by Krastin is erroneous and we illustrate this with an example.

Consider a hypothetical dataset consisting of five meteostations whose climatic conditions have been observed over a period of 3 years. Suppose Y is the annual income in million cedis and X are assets in million cedis, Table 1 presents the data. The regression parameter estimates for each individual year and the classical pooling are presented in Table 2.

Table 1: Hypothetical data

	year 1		year 2		year 3		mean	
Unit	X	Y	X	Y	X	Y	X	Y
1	2	5.2	1	6.2	2	7.4	1.67	6.26
2	4	8.4	2	8.2	3	9.6	3.0	8.73
3	3	6.8	3	10.2	2.5	8.5	4.1	8.5
4	4	8.4	2	8.2	1	5.2	3.67	7.26
5	1	3.6	4	12.2	1.5	6.3	3.03	7.36
Total	14	32.4	12	45	10	37	36	114.4
Mean	2.8	6.48	2.4	9	2	7.4	2.4	7.636

Table 2: Model Estimates

Year	Intercept	Regression Coefficient	Correlation Coefficient
1 2 3	2.0 4.2 3.0	1.6 2.0 2.2	1 1 1
Classical Pooling	3.09	1.549	0.762

One observe that the estimates from the regression model of the classical pooling approach differ paradoxically in magnitude from those of the individual years. Both the regression coefficient and the correlation coefficient of the estimates from the pooled data fall outside the range of the corresponding estimates of the individual years; that is, they are all smaller than the minimum estimates of the individual years. Thus, contrary to Krastin's assertion, the existence of a paradox in classical pooling is also a problem in modelling. Krastin limited himself to the case of two periods and therefore was unable to capture all the components inherent in the equation.

The regression coefficient of the classical pooling can be re-expressed in terms of variances and covariances. In a more compact form we may write it as Eq. (12), where $\bar{x}_{.t}$ and $\bar{y}_{.t}$ are the mean values of the independent and dependent variable, respectively, of the tth year over all units *(t = 1, 2, ..., s, ..., r, ..., T); n = N_t.*

Covariance Structure of the Classical Approach Model

We introduce the concept of covariance (Cov) and variance (s^2) by dividing the various components by n even though we could divide by n – 1. In terms of covariances and variances we rewrite Eq. (12) as Eq. (13) bellow:

$$\widehat{\beta_{cp}} = \frac{\sum_{t=1}^{T}\sum_{i=1}^{n}(x_{it}-\bar{x}_{.t})(y_{it}-\bar{y}_{.t})+\frac{2n}{T^2}\sum_{\substack{s,t=1\\ s\neq t}}^{T}(\bar{x}_{.t}-\bar{x}_{.s})(\bar{y}_{.t}-\bar{y}_{.s})+\frac{n}{T^2}\sum_{\substack{r,s,t=1\\ r\neq s\neq t}}^{T}(\bar{x}_{.t}-\bar{x}_{.s})(\bar{y}_{.t}-\bar{y}_{.r})}{\sum_{t=1}^{T}\sum_{i=1}^{n}(x_{it}-\bar{x}_{.t})^2+\frac{2n}{T^2}\sum_{\substack{s,t=1\\ s\neq t}}^{T}(\bar{x}_{.t}-\bar{x}_{.s})^2+\frac{2n}{T^2}\sum_{\substack{r,s,t=1\\ r\neq s\neq t}}^{T}(\bar{x}_{.t}-\bar{x}_{.s})(\bar{x}_{.t}-\bar{x}_{.r})} \tag{12}$$

$$\widehat{\beta_{cp}} = \frac{\sum_{t=1}^{T} Cov(x_t,y_t)+\frac{2}{T^2}\sum_{\substack{s,t=1\\ s\neq t}}^{T} Cov\left(\bar{x}_s^t,\bar{y}_s^t\right)+\frac{1}{T^2}\sum_{\substack{r,s,t=1\\ r\neq s\neq t}}^{T}(\bar{x}_s^t,\bar{y}_r^t)}{\sum_{t=1}^{T} s^2(x_t)+\frac{2}{T^2}\sum_{\substack{s,t=1\\ s\neq t}}^{T} s^2(\bar{x}_s^t)+\frac{2}{T^2}\sum_{\substack{r,s,t=1\\ r\neq s\neq t}}^{T} Cov(\bar{x}_s^t,\bar{y}_r^t)} \tag{13}$$

where:

$$Cov(x_t, y_t) = \frac{\sum_{i=1}^{n}(x_{it}-\bar{x}_{.T})(y_{it}-\bar{y}_{.T})}{n} \tag{13a}$$

$$Cov(\bar{x}_s^t, \bar{y}_r^t) = \frac{\sum_{i=1}^{n}(\bar{x}_{.t}-\bar{x}_{.s})(\bar{y}_{.t}-\bar{y}_{.r})}{n} \tag{13b}$$

$$Cov(\bar{x}_s^t, \bar{y}_s^t) = \frac{\sum_{i=1}^{n}(\bar{x}_{.t}-\bar{x}_{.s})(\bar{y}_{.t}-\bar{y}_{.s})}{n} \tag{13c}$$

$$s^2(x_t) = \frac{\sum_{i=1}^{n}(x_{it}-\bar{x}_{.t})^2}{n} \tag{13d}$$

$$s^2(\bar{x}_s^t) = \frac{\sum_{i=1}^{n}(\bar{x}_{.t}-\bar{x}_{.s})^2}{n} \tag{13e}$$

$$Cov(\bar{x}_s^t, \bar{x}_r^t) = \frac{\sum_{i=1}^{n}(\bar{x}_{.t}-\bar{x}_{.s})(\bar{x}_{.t}-\bar{x}_{.r})}{n} \tag{13f}$$

Analysis of source of pardox

From the reexpressed formula of the regression coefficient we may obser that both the numerator and the denominator have the components after reexpression.

(a) The numerator contains the following components.

(i) The conventional covariance of *X* and *Y* used in computing the classical regression coefficient; (*Cov*(x_t,y_t)). In addition it also incorporates

(ii) The covariance of the means of *X* and *Y* $Cov(\bar{x}_s^t, \bar{y}_s^t)$ of the *t* and *s* years.

This is the product of the deviation of means of different years of *X* and the deviation of means of corresponding years of *Y*.

(iii) The covariance of the means of *X* and *Y* $Cov(\bar{x}_s^t, \bar{y}_r^t)$ of the *t*,*s* and *r* years.

That is, the product of the means of *X* of two different years *t* and *s* and *Y* of the different years *t* and *r*.

(b) The denominator, or the other hand, contains the following components.

(i) The conventional variance of X used in computing the classical regression co-efficient (*s^2(x_t)).* In addition, it also contains

(ii) The variance of the means of X for two different years ($s^2(\bar{x}_s^t)$).

(iii) The covariance of the means of *X* for different years ($Cov(\bar{x}_s^t, \bar{x}_r^t)$).

That is, the product of the means of *X* of two different years *t* and *s* that of *X* of the different years *t* and *r*.

Thus, the classical pooling generates a formula for computing regression coefficient which incorporates some terms (a(ii), a(iii), b(ii), b(iii)) that are not in the conventional formula. Krastin ([Kra81]) using only two years in his work identified only two, that is, those of a(ii) and b(ii).

3. CONCLUSION

Model validation is the final step in the model-building process and it usually involves checking of the model against independent data or whether the model will function successfully in its intended operating environment. Validation requires that the model prediction performance be investigated (by considering both interpolation and extrapolation) to check the stability of the model. Validation gives the assurance that the model developed primarily for prediction provides the fit to the existing data. Hence the model to be used must be appropriate so as to achieve its goal. Therefore, the study shows ANCOVA (analysis of covariance) as the appropriate technique for handling big data. ANCOVA contains a mixture of qualitative variables associated with analysis of variance and the quantitative variables associated with regression analysis. ANCOVA is the meeting point under the umbrella of analysis of variance and regression techniques.

REFERENCES

[BC11] **D. Boyd, K. Crawford** - *Six Provocations for Big Data*. Social Science Research Network: A Decade in Internet Time: Symposium on the Dynamics of the Internet and Society. *doi:10.2139/ssrn.1926431*, 2011.

[C+16] **S. Cheng, B. Liu, T. O. Ting, Q. Qin, Y. Shi, K. Huang** - *Survey on data science with population-based algorithms*, DOI 10.1186/s41044-016-0003-3. 2016.

[ES07] **R. Eberhart, Y. Shi** - *Computational Intelligence: Concepts to Implementations*. San Francisco: Morgan Kaufmann Publisher; 2007.

[Gre93] **D. M. Greene** - *Econometric Analysis*, (2nd ed.). New York: Macmillan, 1993.

[Gri16] **S. Grimes** - *Big Data: Avoid 'Wanna V' Confusion*, Information Week. Retrieved 5 January 2016.

[Hil15] **M. Hilbert** - *Big Data for Development: A Review of Promises and Challenges. Development Policy Review*, http://martinhilbert.net, retrieved 7 October 2015.

[H+15] **I. A. T. Hashem, I. Yaqoob, N. B. Anuar, S. Mokhtar, A. Gani, S. U. Khan** - *The rise of "big data" on cloud computing: Review and open research issues*. Information Systems. **47**: 98–115. doi:10.1016/j.is.2014.07.006. 2015.

[Kra81] **O. P. Krastin** - *Vaprosi premineniya srednix mnogoletnik dannix vo regressionnom analisie*, Vestnik Statistiki, 7.1981.

[Nua86] **N. N. N. N. Nuamah** - *Pooling cross section and time series data*, The Statistician, 35, 345-351, 1986.

[Nua92] **N. N. N. Nuamah** - *The Error and Covariance Structures of the Mean Approach of Pooled Cross-Section and Time Series Data*, Statistician, 41, 197-207, 1992.

[Nua00] **N. N. N. Nuamah** - *Classical Pooling of Cross-Section and Time Series Data*, ICTP IC/99/169, Trieste, Italy, 2000.

[W+14] **X. Wu, X. Zhu, G. Q. Wu, W. Ding** - *Data mining with big data*. IEEE Trans Knowl Data Eng. 2014;26(1):97–107.

[***18] *** - *What is Big Data?*. https://www.villanovau.com/resources/bi/what-is-big-data/, Retrieved 2018.