NAVIGATING THE DATA LABYRINTH: Applications of Some Advanced Statistical Analysis in Atmospheric Physics

BAYERO UNIVERSITY KANO
PROFESSORIAL INAUGURAL LECTURE
NO. 33

Bello Idrith Tijjani
Physics Department,
Faculty of Physical Sciences,
Bayero University, Kano

DATE: 20th September 2018

Published 2018 by:
Bayero University Press,
Main Library Building,
Bayero University Kano,
New Site, Gwarzo Road,
P.M.B. 3011.
Kano.

Website: *www.buk.edu.ng*
E-mail: *info.bukpress@buk.edu.ng*

ISBN 978 - 978 - 56076 - 8 - 0

PRINTED BY BAYERO UNIVERSITY PRESS KANO, NIGERIA.

Bello Idrith Tijjani

B Sc. M.Sc. PhD (Physics), M. Eng. (Electrical)(BUK) MNAMP, MNIP
Professor of Atmospheric Physics,
Physics Department, Faculty of Physical Sciences,
Bayero University, Kano.

SUMMARY OF PRESENTER'S BIODATA

Professor B. I. Tijjani, *B.Sc. M.Sc. PhD (Physics), M. Eng. (Electrical)(BUK) MNAMP, MNIP*, is a Professor of Atmospheric Physics in the Physics Department of Bayero University, Kano.

Born on 16th August 1962 at Kabo in Kabo Local Government area of Kano State, Professor Bello Idrith Tijjani attended Massallachi Primary School in Kano Municipal from 1969 to 1976. Upon completing his primary education, he proceeded to Government Secondary School Kazaure, where he spent three years, from 1976 to 1978. From 1978 to 1981, he was at Science Secondary School Dawakin Kudu, Kano State where he finished his secondary school education leading to his G.C.E O'Level certificate. The following year, he secured admission for pre-degree at Bayero University, Kano, and after completing the course he proceeded for his B.Sc. degree in Physics which he completed in 1987. In 1989, he enrolled for his Master of Science programme in the same Department and completed it in 1992. Professor Bello Idrith Tijjani commenced his PhD programme in Theoretical Physics in the year 1993 and finished in 1999. Still further propelled by the desire to broaden his intellectual frontiers, he enrolled for another Masters degree in Electrical Engineering (M.Eng. (Electrical)) between 2001 and 2014.

Professor Bello Idrith Tijjani did the compulsory National Youth Service at Hassan Usman Katsina Polytechnic (formerly Katsina Polytechnic) from 1987 to 1988. He joined the services of Bayero University as a Graduate Assistant in the Department of Physics of the then Faculty of Science from 1988 to 2001 before he was transferred to the Department of Mathematical Sciences from 2001 to 2009, where he contributed to the establishment of the Computer Science programme. In 2009, he came back to Physics Department to further his career in Physics and has been there till date. He rose through the ranks and became a professor in 2016.

From 1988 to date, he has taught several courses in Physics and Computer Science at both the undergraduate and postgraduate levels. His research in Physics are: Theoretical and Computational Physics with applications to atmospheric modelling, radiative transfer, effect of particle size on hygroscopic behaviour and on deliquescence and efflorescence, the effects of relative humidity on aerosol chemical, physical, and optical properties, phase transitions of aerosol particles; Light scattering and radiative transport in the atmosphere. The research areas in Computer Science are: Data Communications, Network Modelling.

B.I.T, as he is fondly called, is quite very conversant with FORTRAN, VISUAL BASIC and C++ programming languages, and SPSS, MATLAB and EXCEL application packages for statistical analysis and modelling.

In the course of conducting his teaching and research activities, he has fully authored and published the following textbooks:

- *Introduction to Vectors and Tensors, with Applications to Physics" (2013)*
- *A Guide to FORTRAN Programming (2002)*
- *Introduction to VISUAL C++ Programming (2003)*
- *A Guide to FORTRAN Programming (2007)*

His unpublished works include: "Introduction to Electromagnetism with Relativity" and "Introduction to Finite Elements with applications to Physics". He also supervised forty three (43) M.Sc. Computer Science students, fourteen (14) M.Sc. Physics students, and five (5) PhD Physics students. He also served as an internal examiner to many M.Sc. and PhD students in the University. In addition, he has to his credit, over forty (40) journal articles published in various journals both locally and internationally.

Professor Bello Idrith Tijjani has attended several conferences and workshops organized by different institutions at different levels. He has served as an external moderator of question papers for Physics and Computer Science for different colleges of education and universities. He is a visiting lecturer to Ahmadu Bello University, Zaria; Yusif Maitama Sule University, Kano; Kano University of Science and Technology, Wudil, Kano; Umaru Musa Yar'Adua University, Katsina; Gombe State University and Alqalam University, Katsina.

An active member of the University community, BIT has at different times served as level coordinator at both undergraduate and postgraduate levels. He also served as Sub-Dean from 1995 to 1998 among different numerous responsibilities.

NAVIGATING THE DATA LABYRINTH:

Applications of Some Advanced Statistical Analysis in Atmospheric Physics

PREAMBLE

It is a great honour and privilege for me to be able to present for academic digestion, some of the elements of the discipline that fascinates me. I am, therefore, immensely grateful to the Almighty Allah for the opportunity to stand before this audience to give this lecture.

INTRODUCTION

Computers have changed the way statistics is learned and taught. Often, researchers are interested only in the “results” of their “analyses” and do not care about how the results are obtained. With the advancement in the field of information and communication technologies, it has become easier to capture huge amounts of data. However, the sheer amount of data makes it virtually impossible to comprehend them in their raw form. The purpose of this presentation is to discuss some applications of some advanced statistics analysis that are very useful in the analysis of data in atmospheric physics.

Some descriptive statistics that are normally used to summarize and present data in a meaningful manner so that the underlying information is easily understood are discussed in section 2. In the section mean, median, mode, skewness and kurtosis are discussed. The last two are discussed in a greater detail due to their importance. This is attributable to the fact that they are not commonly used and even if used, are not normally properly interpreted.

In Section 3, time series analyses are discussed. The reason for discussing it is because the way many textbooks give guides on how to use them makes it very difficult for a non-statistician to understand. Some of the steps include: how to check the data for test of stationarity using AutoCorrelation Function (ACF) and Partial AutoCorrelation Function (PACF) and how to take seasonal differencing and non-seasonal differencing. In this section, some meteorological data are analysed using the Expert Modeler of SPSS 16.0, although SPSS 20.0 IBM version also has the Expert Modeler. The advantage of this approach is that the software optimizes all optimizable parameters and finally gives the most appropriate model(s) as it displays the model(s) as either, ARIMA or Exponential Smoothing Model.

In Section 4, analyses are made on how to use Empirical Orthogonal Functions (EOFs). In this case the data are converted to matrix form, and when inputted in the software, the correlation/covariance matrix is obtained. The software uses the correlation/covariance matrix to determine the eigenvalues and eigenvectors. Usually, these values represent the types and number of modes of any system. The metrological data used in section 3 is used here and SPSS 16.0 is also used, although SPSS 20.0 IBM version can also be used.

Finally, the focal point of this presentation is on how SPSS is used as a black box. In the process, the discussions of the theories are limited to the necessary and simple terms of significance. The manner by which outputs are interpreted using examples is also made evident.

UNDERSTANDING AND INTERPRETING PARTICLE SIZE DISTRIBUTION USING DESCRIPTIVE STATISTICS

Introduction

It is typically noted in introductory statistics courses that distributions can be characterized in terms of central tendency, variability, and shape. In atmospheric physics, performing a particle size analysis is the best way to answer the question: What size are those particles? Once the analysis is complete the user has a variety of approaches for reporting the results. The need to study these concepts arises from the fact that the measures of central tendency and dispersion alone fail to describe a distribution completely. It is possible to have frequency distributions which differ widely in their nature and composition and yet may have the same central tendency and dispersion. Thus, there is the need to supplement the measures of central tendency and dispersion. Some of the supplements to be discussed are the numerical methods of the measures of shapes (skewness and kurtosis).

The purpose of this section is to clarify the meanings of kurtosis and skewness and to show why and how they can be used in atmospheric physics. The link between skewness and kurtosis, and Angstrom constants are been examined and the necessity of their joint use is being justified.

MEASURES OF CENTRAL TENDENCY

Measures of central tendency provide information about a representative value of the data set. Arithmetic mean, simply called the mean, the median, and the mode are the most common measures of central tendency.

(i) Mean or average is the sum of the values of a variable divided by the number of observations.

(ii) Median is a point in the data set above and below which half of the cases fall. Median values are defined as the value where half of the population resides above this point, and half resides below this point.

(iii) Mode is the most frequently occurring value in a data set. The mode is the peak of the frequency distribution, or it may be easier to visualize it as the highest peak seen in the distribution. The mode represents the particle size (or size range) most commonly found in the distribution.

For symmetric distributions, such as the one shown in Figure 2.1, all central values are equivalent: mean = median = mode. The importance of these parameters in atmospheric physics is what is going to be presented. For non-symmetric distributions the mean, the median and the mode will be the three different values shown in Figure 2.1.

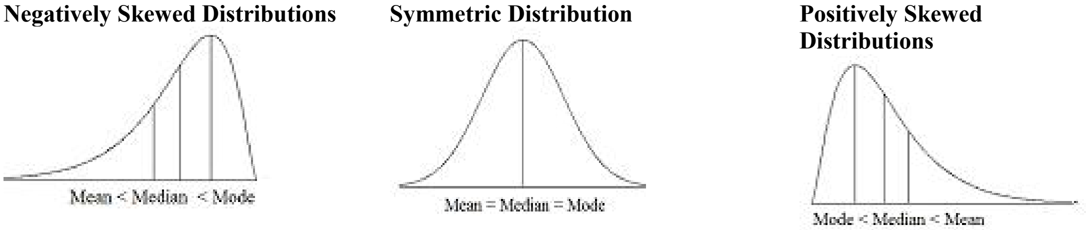
MEASURES OF SHAPES: SKEWNESS AND KURTOSIS

A histogram can give a general idea of the shape of a distribution, but two numerical measures of shape give a more precise evaluation: skewness tells you the amount and direction of skew (departure from horizontal symmetry), and kurtosis tells you how tall and sharp the central peak is, relative to a standard bell curve (vertical measures).

Skewness

Besides mean, median, and mode, it is also important to know if the given distribution is symmetric or not. A distribution is said to be skewed if the observations above and below the mean are not symmetrically distributed. A zero value of skewness implies a symmetric distribution. The distribution is positively skewed when the mean is greater than the median and negatively skewed when the mean is less than the median. A textbook discussing the concepts would typically begin by showing the relative positions of the mean, median, and mode in smooth population probability density functions, as illustrated in Figure 2.1.

Figure 2.2 shows how histograms can be used to give the general idea of the different types of size distributions.

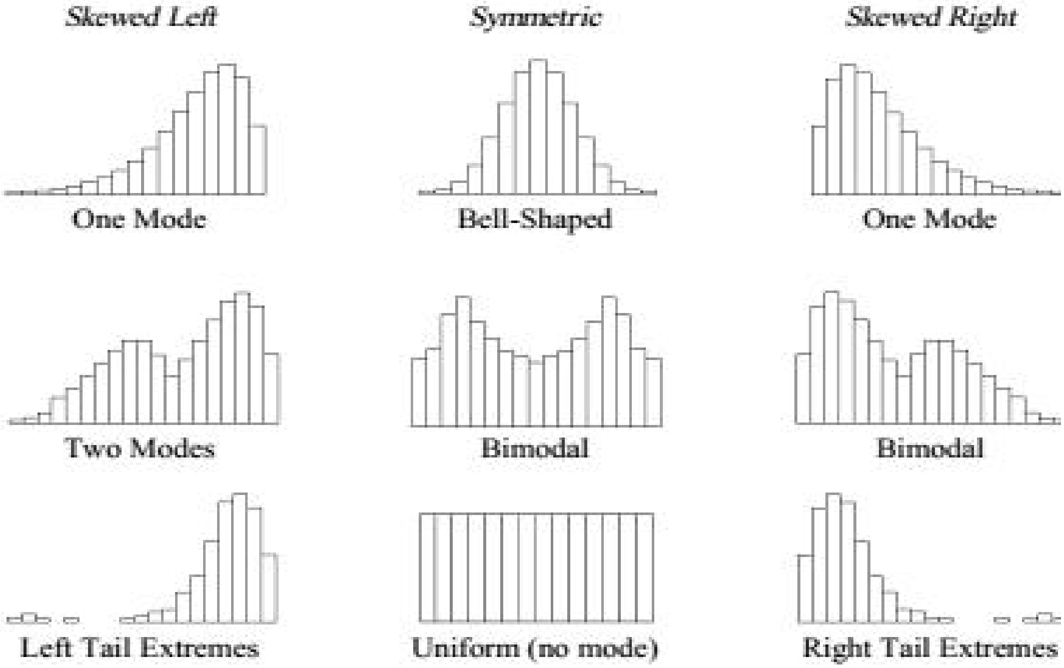


Negatively Skewed Distributions

Symmetric Distribution

Positively Skewed Distributions

Figure 2.1: *Sketches showing general position of mean, median and mode in a population.*



Skewed Left

Symmetric

Skewed Right

Figure 2.2: *Illustrative prototype histograms*

From Figures 2.1 and 2.2, it is observable that for negatively skewed (Skewed left), there are long tails and distortions that are caused by extremely small values which pull the mean downward so that it is less than the median. At the centre, there are symmetric distributions in which the mean, the median and the mode are the equal. By the right hand side of the figures there are long tails to the right caused by extremely large values which pull the mean upward so that it is greater than the median.

Table 2.1: *Description of different types of skewness*

Skewness	Distribution shape	Calculated Value
Positive	Tail to the right, values extend further to the right but concentrated in the left	Mean>Median>Mode
Zero	Bell shaped or symmetrical	Mean=Median=Mode
Negative	Tail to the left, values extend further to the left but concentrated in the right.	Mean<Median<Mode

Kurtosis

If a distribution is symmetric, the next question is about the central peak: Is it high and sharp, or short and broad? You can get some idea of this from the histogram, but a numerical measure is more precise. Kurtosis is a measure of how peaked or flat a distribution is or is the degree of peakedness of a distribution, that is, whether it is of peaked or flat relative to a normal distribution (its departure from the vertical with respect to the normal distribution). The reference standard is a normal distribution, which has a kurtosis of 3. Often, excess kurtosis is presented instead of kurtosis, where excess kurtosis is simply “kurtosis – 3”. For example, the “kurtosis” reported by Excel and SPSS is actually the excess kurtosis.

Data sets with high kurtosis tend to have a distinct peak near the mean, decline rather rapidly, and have heavy tails. Data sets with low kurtosis tend to have a flat top near the mean rather than a sharp peak.

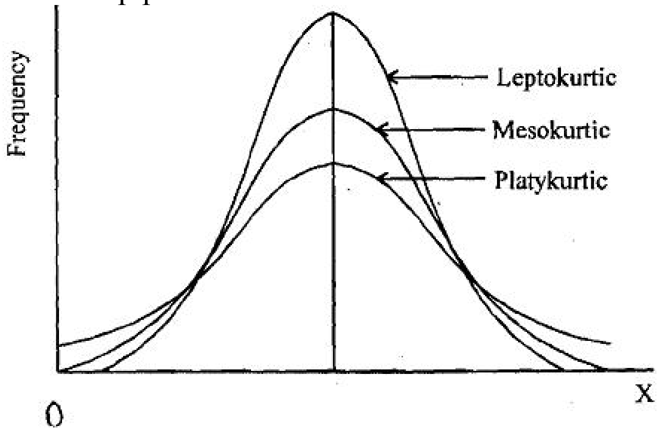


Figure 2.3: *Meso, Lepto, and Platykurtic Distributions*

Table 2.2: *Description of different types of kurtosis*

Term	Distribution shape	Kurtosis	Excess Kurtosis
Leptokurtic	Peaked	Greater than 3	Greater than 0
Mesokurtic	Normal	3	0
Platykurtic	Flat	Less than 3	Less than 0

EXAMPLE: ANALYSIS OF ATMOSPHERIC AEROSOLS SIZE DISTRIBUTIONS USING SKEWNESS, KURTOSIS AND ANGSTROM CONSTANTS

Introduction

Aerosol particles result from different sources and processes. At any place in the atmosphere there exists a mixture of particles of different origin. To describe the wide range of possible compositions, the aerosol particles are modelled as components (Deepak and Gerber, 1983), each of them meant to be representative for a certain origin, that is, an internal mixture of all chemical substances that have a similar origin. These components may be externally mixed to form aerosol types. External mixture means that there is no physical or chemical interaction between particles of different components.

The aim of this presentation is to compare skewness and kurtosis with Angstrom exponents (Angstrom's turbidity coefficient β and the wavelength exponent α) in the analysis of the atmospheric aerosols particles size distributions. In atmospheric physics, Angstrom exponents are used as measures to determine the aerosols size distributions. Higher values, usually greater than 1, represents the dominance of fine/accumulation modes over coarse modes particles, while lower down to negative indicates the dominance of coarse mode particles over fine/accumulated mode particles. The atmospheric aerosols to be used are desert and urban aerosols. These aerosols are extracted from Optical Properties of Aerosols and Clouds (OPAC) 4.0, at the relative humidities (RHs) of 0, 50, 70, 80, 90, 95, 98, and 99% at wavelengths of 0.40 to 0.8μm.

Methodology

The aerosols models types extracted from OPAC 4.0, at the relative humidities of 0, 50, 70, 80, 90, 95, 98, and 99% at wavelengths of 0.4 to 0.8μm are Desert and Urban aerosols. The software mixes the components externally to form the aerosols types.

Table 2.3: *Compositions of aerosols types at 0% RH (Hess et. al., 1998).*

Aerosols Models	**Aerosols Components**	**Number Concent. (cm^{-3})**	**Number Mix Ratio**	**Volume Mix Ratio**	**R$_{mod}$ (dry), (µm):**
Desert	waso	2000	0.8695	0.01842	0.0212
	minn	269.5	0.1172	0.03474	0.0700
	mian	30.5	0.01326	0.7442	0.3900
	micn	0.142	0.00006174	0.2026	1.9000
Urban	inso	1.5	0.000009494	0.3832	0.4710
	waso	28000	0.1772	0.4493	0.0212
	soot	130000	0.8228	0.1675	0.0118

The *inso* represents the *water-insoluble* part of aerosol particles and consists mostly of soil particles with a certain amount of organic material. The *waso* represents the *water-soluble* part of aerosol particles that originates from gas to particle conversion and consists of various kinds of sulfates, nitrates, and other, also organic, water-soluble substances. Thus, it contains more than only the sulfate aerosol that is often used to describe anthropogenic aerosol. The *soot* component is used to represent absorbing black carbon. Carbon is not soluble in water and, therefore, the particles are assumed not to grow with increasing relative humidity. *Mineral* aerosol or desert dust is produced in arid regions. It consists of a mixture of quartz and clay minerals and is modelled with three modes to allow and consider increasing relative amount of large particles for increasing turbidity. mineral (nuclei mode, nonspherical) *minn*, mineral (accumulation mode, nonspherical) *mian*, mineral (coarse mode, nonspherical) *micn* are mineral aerosols or desert dusts that are produced in arid regions.

With rising humidity, the aerosol particles are more and more soaked with water from the surrounding humid air and swell. The increase in particle size reduces the visibility. Quantitatively, the variation of the size distribution of the aerosol particles with relative humidity has to be taken into account (Kasten, 1968). An objective measure of visibility is the standard visual range or meteorological range (Koschmieder, 1926)

$$V(\lambda) = \frac{3.912}{\sigma_{ext}(\lambda)} \tag{2.1}$$

which is meteorological range refers to the visual range of a black object seen against the horizon sky by a standard observer having a contrast threshold 0.02 (Middleton, 1952). The visual extinction coefficient $\sigma_{ext}(\lambda)$ is a measure of the light scattering and absorbing properties of the atmosphere along the line of sight.

Variation of the extinction coefficient with wavelength can be presented in the form of an inverse power law function (Angstrom, 1929); that is the spectral dependence of extinction by particles may be approximated as an inverse power law relationship:

$$\sigma_{ext}(\lambda) = \beta\lambda^{-\alpha} \qquad (2.2)$$

where α and β are known as Angstrom parameters . The index α is the wavelength exponent or Angstrom coefficient; β is the turbidity coefficient representing the amount of aerosols present in the atmosphere in the vertical direction or the total aerosol loading in the atmosphere (Shaw *et. al.,* 1973; Satheesh and Moorthy, 1997). The Angstrom exponent depends on the size distribution of aerosols and is a measure of the ratio of the concentration of coarse to accumulation mode aerosols, with higher values representing increased abundance of accumulation mode aerosols. Higher and positive values of α indicate dominance of fine/accumulation mode aerosols in the aerosol size spectrum, whereas lower and negative values of α indicate the dominance of coarse mode aerosol particles (Moorthy *et. al.,* 2001; Singh *et. al.*, 2005).

One of the most important variables for the aerosol size distribution is the effective radius, which can be calculated using the Junge size distribution n(r) (Junge, 1963):

$$n(r) = \frac{dN(r)}{dr} = C(z)r^{-\,(\nu+1)} \qquad (2.3)$$

where N is the number density, r is the radius and C(z) is a factor proportional to the aerosol concentration, which is dependent on the altitude z (Biggar *et. al.*, 1990). Through the relationship ν=α+2(Iqbal, 1983; Bruegge *et. al.,* 1992), the exponent ν can be obtained from the value of α, which can be estimated from equation (2.2).

In terms of individuals aerosols size distributions, lognormal distributions (cf.,e.g., Deepak and Gerber, 1983) are applied for each component i:

$$\frac{dN_i(r)}{dr} = \frac{N_i(r)}{\sqrt{2\pi} r log\sigma_i ln10} exp\left[\frac{1}{2}\left(\frac{logr - logr_{modN_i(r)}}{log\sigma_i}\right)^2\right] \qquad (2.4)$$

where $r_{modN,i}$ is the mode radius, σ_i measures the width of the distribution, and N_i is the total particle number density of the component *i* in particles per cubic centimetre.

Substituting equation (2.2) into (2.1), the following equation is obtained also as a direct power law function

$$V(\lambda) = \frac{3.912}{\beta}\lambda^{\alpha} \qquad (2.5)$$

Equation (2.5) can also be written as:

$$ln\left(\frac{V_\lambda}{3.912}\right) =- ln(\beta) + \alpha\, ln(\lambda) \qquad (2.6)$$

Measurements indicate that the Angstrom exponent varies with wavelength, and a more precise empirical relationship between aerosol extinction and wavelength is obtained with a 2nd-order polynomial (King and Byrne, 1976; Eck *et. al.,* 1999; Eck. *et. al.,* 2001a, b; Kaufman , 1993; O'Neill et al., 2001, 2003; Galadanci and Tijjani, 2014; Tijjani *et. al.,* 2013; Tijjani *et. al.,* 2014). This implies that it can also be applicable to visibility:

$$ln\left(\frac{V_{\lambda}}{3.912}\right) =- ln(\beta) + \alpha_1 ln(\lambda) + \alpha_2(ln(\lambda))^2 \qquad (2.7)$$

Here, the coefficient α_2 accounts for a" curvature" often observed in sunphotometry measurements. The spectral curvature of the Angstrom exponent contains useful information about the aerosol size distribution (King and Byrne, 1976; Eck *et. al.,* 1999; Eck. *et. al.,* 2001a, b; Kaufman, 1993; O'Neill *et. al*., 2001, 2003). Some authors have noted that the curvature is also an additional indicator of the aerosol particle size, with negative curvature indicating aerosol size distributions dominated by the fine mode or monomodel distribution and positive curvature indicating size distributions with a significant coarse mode contribution or bimodal size distribution (Kaufman, 1993; Eck *et. al*., 1999; Eck. *et. al.,* 2001b).

Results and Discussions

The results, the analysis and observations of the data extracted from OPAC 4.0, is now presented.

Desert

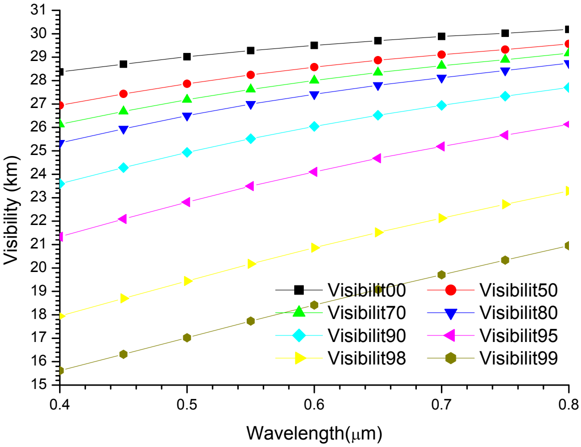


Figure 2.4: *The plots of visibilities with wavelengths for Desert aerosols*

From the plots of figure 2.4, it can be seen that the visibility increases with the increase in wavelength, but decreases with the increase in RH.

Table 2.2: Results of the regression analysis of equation (2.6) for Desert aerosols

RH (%)	R^2	Sig.	α	Sig.	β	Sig.
0	0.9963	8.87E-10	0.0892	8.87E-10	0.0324	1.59E-22
50	0.9965	7.43E-10	0.1332	7.43E-10	0.0328	2.25E-21
70	0.9970	4.22E-10	0.1580	4.22E-10	0.0330	4.27E-21
80	0.9974	2.59E-10	0.1811	2.59E-10	0.0333	6.95E-21
90	0.9988	1.80E-11	0.2323	1.80E-11	0.0342	2.89E-21
95	0.9995	5.75E-13	0.2941	5.75E-13	0.0358	5.29E-22
98	0.9998	2.44E-14	0.3785	2.44E-14	0.0395	1.62E-22
99	0.9991	7.63E-12	0.4269	7.63E-12	0.0436	1.46E-19

Based on the contents of Table 2.2 and by observing the values of R^2 and the significances of all the coefficients, it can be said that the data fitted the equation model very well (equation 2.6). The increase of α with RH signifies the increase in the dominance of fine/accumulation modes over coarse modes particles. But since it is less than 1 it shows that coarse mode particles are still more dominant. And β (turbidity coefficient) increases with the increase in RH, and this also contributes to the decrease in visibility with RH.

Table 2.3: *Results of the regression analysis of equation (2.7) for Desert aerosols*

RH (%)	R^2	Sig.	α_1	Sig.	α_2	Sig.	β	Sig.
0	0.9998	5.96E-12	0.0590	8.17E-07	-0.0270	3.78E-05	0.0326	6.03E-21
50	0.9998	3.71E-12	0.0891	4.78E-07	-0.0391	2.74E-05	0.0331	4.26E-20
70	0.9999	1.15E-12	0.1096	1.21E-07	-0.0429	1.38E-05	0.0334	3.73E-20
80	0.9999	1.42E-12	0.1296	1.23E-07	-0.0457	2.58E-05	0.0337	1.06E-19
90	0.9999	2.48E-13	0.1874	1.06E-08	-0.0399	4.45E-05	0.0345	8.58E-20
95	0.9999	5.61E-13	0.2618	1.32E-08	-0.0286	0.00203	0.0360	8.60E-19
98	0.9999	5.20E-13	0.4004	4.35E-09	0.0195	0.03222	0.0393	4.23E-18
99	0.9999	3.08E-13	0.4992	1.42E-09	0.0641	1.16E-04	0.0428	6.09E-18

Considering Table 2.3 and by observing the values of R^2 and the significances of all the coefficients, it can be said that the data fitted the equation model very well. The increase of α_2 with RH signifies the increase in the dominance of fine/accumulation modes over coarse modes particles. The negative sign of α_2, shows that this is monomodal and is dominated by coarse mode particles (at RH of 0 to 95%). But at 98 and 99% RH, it becomes positive. This is where the bimodal distribution appears. But

β (turbidity coefficient) increases with the increase in RH, and this also contributes to the decrease in visibility with RH.

Table 2.4: The skewness and kurtosis of Desert aerosols

	Vis00	Vis50	Vis70	Vis80	Vis90	Vis95	Vis98	Vis99
Mean	29.408	28.437	27.854	27.255	25.876	23.945	20.753	18.354
Median	29.502	28.576	28.003	27.414	26.045	24.104	20.864	18.418
Mode	28.368	26.942	26.132	25.337	23.595	21.330	17.945	15.617
Skewness	-0.476	-0.458	-0.437	-0.417	-0.350	-0.284	-0.165	-0.088
Kurtosis	-0.878	-0.889	-0.922	-0.939	-1.042	-1.106	-1.187	-1.210

As indicated by Table 2.4 and by observing the decrease in the values of mean, median and mode, it can be said that there is a decrease in the number of particles as the relative humidity increased. By observing the skewness, it can be seen that they are all negative (negatively skewed), this is an indication of the dominance of coarse mode particles compared to fine mode particles. The decrease in magnitude with RH shows that larger particles are decreasing more in number than the fine particles. From the kurtosis, it can be said that it is negative (platykurtic), and this shows that the distribution is below the normal distribution. The increase in the magnitude with the increase in RH reflects the decrease in the particles as they are removed from the atmosphere.

Comparing α and the skewness they both show the dominance of coarse mode particles. This is because α is less than 1.0 and the skewness is negative. From the sign of α_2 (negative), it shows that the aerosols have monomodel distribution, and the signs of the kurtosis (platykurtic) show the dominance of coarse mode particle distribution.

Urban

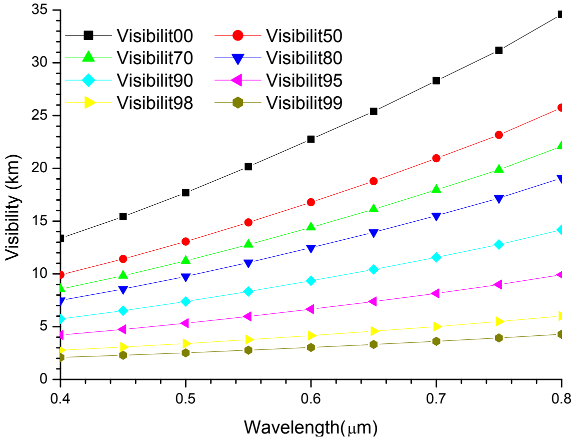


Figure 2.5: *The plots of Visibilities with wavelengths for Urban aerosols*

From the plots of figure 2.5, it can be seen that the visibility increases with the increase in wavelength, but decreases with the increase in RH.

Table 2.5: *Results of the regression analysis of equation (2.6) for Urban aerosols*

RH (%)	**R^2**	**Sig.**	**α**	**Sig.**	**β**	**Sig.**
0	0.9990	9.17E-12	1.373	9.17E-12	0.0216	1.52E-16
50	0.9984	4.33E-11	1.379	4.33E-11	0.0291	1.31E-15
70	0.9982	7.71E-11	1.369	7.71E-11	0.0341	3.05E-15
80	0.9978	1.39E-10	1.353	1.39E-10	0.0396	6.97E-15
90	0.9972	3.32E-10	1.309	3.32E-10	0.0540	2.68E-14
95	0.9964	8.43E-10	1.239	8.43E-10	0.0785	1.21E-13
98	0.9951	2.47E-09	1.123	2.47E-09	0.1332	9.09E-13
99	0.9939	5.07E-09	1.036	5.07E-09	0.1906	4.22E-12

Through the keen observation of table 2.5 and by observing the values of R^2 and the significances of all the coefficients, it can be said that the data fitted the equation model very well. The increase of α with RH signifies the increase in the dominance of fine/accumulation modes over coarse modes particles. But β (turbidity coefficient) increases with the increase in RH, and this also contributes to the decrease in visibility with RH.

Table 2.6: *Results of the regression analysis of equation (2.7) for Urban aerosols*

RH (%)	R2	Sig.	α_1	Sig.	α_2	Sig.	β	Sig.
0	0.9999	1.62E-13	1.6137	7.24E-10	0.2136	5.18E-05	0.0204	9.98E-16
50	1.0000	5.16E-14	1.6856	1.83E-10	0.2721	4.31E-06	0.0271	5.13E-16
70	1.0000	3.96E-14	1.7011	1.27E-10	0.2943	2.01E-06	0.0315	4.88E-16
80	1.0000	3.37E-14	1.7104	9.79E-11	0.3170	1.03E-06	0.0364	5.01E-16
90	1.0000	2.32E-14	1.7010	5.70E-11	0.3480	3.36E-07	0.0492	5.01E-16
95	1.0000	1.12E-14	1.6649	2.26E-11	0.3775	7.29E-08	0.0710	3.80E-16
98	1.0000	7.50E-15	1.5729	1.19E-11	0.3993	1.95E-08	0.1196	5.30E-16
99	1.0000	7.41E-15	1.4975	9.77E-12	0.4090	1.04E-08	0.1708	9.78E-16

From Table 2.6, by observing the values of R^2 and the significances of all the coefficients, it can be said that the data fitted the equation model very well. The increase of α_1 with RH (0 to 80%) signifies the increase in the dominance of fine/accumulation modes over coarse modes particles. But its decrease from 80 to 99%, signifies increase in the concentrations of coarse mode particles as the fine particles are acting as cloud condensation nuclei. From the sign of α_2, it can be seen that it is positive and this shows that it is a bimodal type of particle distributions with the dominance of fine mode particles. The increase in the values of $\alpha_{2,}$ with RH shows the increase in the concentrations of coarse as the fine particles are becoming bigger as they act as cloud condensation nuclei. But β (turbidity coefficient) increases with the increase in RH and this also contributes to the decrease in visibility with RH.

Table 2.7: *The skewness and kurtosiss of Urban aerosols*

	Vis00	Vis50	Vis70	Vis80	Vis90	Vis95	Vis98	Vis99
Mean	23.207	17.190	14.760	12.790	9.583	6.815	4.249	3.095
Median	22.757	16.790	14.404	12.474	9.348	6.651	4.157	3.033
Mode	13.365	9.911	8.560	7.483	5.730	4.211	2.767	2.093
Skewness	0.211	0.240	0.248	0.256	0.261	0.264	0.259	0.256
Kurtosis	-1.134	-1.116	-1.111	-1.105	-1.102	-1.103	-1.108	-1.115

Based on Table 2.7 and by observing the decrease in the values of mean, median and mode, it can be said that there is a decrease in the number of particles as the relative humidity increases. By observing the skewness, it can be said that it is positively skewed. This shows that there is a dominance of fine mode particles. Its increase with the increase in RH from 0 to 95% shows that finer particles are becoming more in number compared to coarse particles. But from 95 to 99% there is a decrease, and this shows the position where the soot particles became cloud condensation nuclei. From the perspective of the kurtosis, it is negative (platykurtic). This shows that the particles are below normal; this is due to the dominance of the fine mode particles as observed in the skewness. The decrease in the magnitude of the kurtosis with the increase in RH (0 to 90%) shows that there is an increase in the concentration of fine particles. But the increase from 90 to 99% shows that some of the larger have started sedimenting and the fine particles have started becoming bigger by acting as cloud condensation nuclei.

By comparing α and skewness they both show the dominance of fine mode particles. This is because, α is greater than 1.0 and the skewness is positive. The signs of α_2 (positive) show the presence of coarse mode particles. Therefore, this shows that the particle distribution is bimodal with the dominance of fine mode particles.

Conclusion

From the analysis of the data, it can be concluded that skewness can be used for verification to determine the most dominant types of aerosols after determining the Angstrom exponents. This is because, there some arguments on the range of values of the exponents to ascertain whether the distribution is fine mode or coarse mode. Now, α_2 can be used to determine model distribution whether monomodal or bimodal. This shows that there is the need to modify the importance of curvature as pointed out by some researchers (Kaufman, 1993; Eck *et. al*., 1999; Eck. *et. al.,* 2001b) on its use as an additional indicator of the aerosol particle size with negative curvature indicating aerosol size distributions dominated by the fine mode or monomodel distribution and positive curvature indicating size distributions with a significant coarse mode contribution or bimodal size distribution. From the kurtosis perspective, it can be used to determine the degree of the variabilities between the average particle sizes and the most common (mode).

TIME SERIES ANALYSIS

Introduction

A time series (TS) is a time-oriented or chronological sequence of observations on a variable of interest (Montgomery et al., 2008). Mostly, these observations are collected at equally spaced, discrete time intervals. When there is only one variable upon which observations are made, such is called a single time series or, more specifically, a univariate time series. Time series models have become popular in recent years since the publication of a book by Box and Jenkins (1970), and the subsequent development of computer software for applying these models (Bell, 1984).

Time series data arise in virtually every application field, such as:

(i) Business: sales figures, production numbers, customer frequencies, ...
(ii) Economics: Stock prices, exchange rates, interest rates, ...
(iii) Official Statistics: Census data, personal expenditures, road casualties, ...
(iv) Natural Sciences: Population sizes, sunspot activity, chemical process data, ...
(v) Environmetrics: Precipitation, temperature or pollution recordings, ...

A basic assumption in any time series analysis modelling is that some aspects of the past pattern will continue to remain in the future. The objective of time series analysis is generally to understand and identify the stochastic process that produced the observed series and then to forecast future values of a series from past values alone (Akgun, 2003).

In this presentation, our goal is to promote the intuitive understanding of seemingly complicated time series models and their implications. We employ only the necessary amount of theory and attempt to present major concepts in time series analysis via an example.

THEORY

While the theory on mathematically oriented time series analysis is vast and mostly difficult to non-statistician, the focus of this presentation is directed at data analysis. Some basic properties of time series processes and models are presented. These focus mostly on how to visualize and describe time series data, on how to fit models to data correctly, on how to generate forecasts and on how to adequately draw conclusions from the output that was produced.

Time Series Components and Decomposition

An important step in analysing TS data is to consider types of data patterns, so that the models most appropriate to those patterns can be utilized. Four types of time series components can be distinguished. They are:

(i) Horizontal – when data values fluctuate around a constant value.
(ii) Trend – when there is a long term increase or decrease in the data.
(iii) Seasonal – when a series is influenced by season factor and reoccurs on regular periodic basis.
(iv) Cyclic – when the data exhibit rise and fall that are not of a fixed period.

Many data series include combinations of the preceding patterns. After separating the existing pattern in any time series data, the pattern that remains unidentifiable form the ‘random’ or ‘error’ component. Time plot (data plotted over time) and seasonal plot (data plotted against individual seasons in which the data were observed) help in visualizing these patterns while exploring the data. A crude, yet practical, way of decomposing the original data (ignoring cyclic pattern) is to go for a seasonal decomposition either by assuming an additive or multiplicative model viz:

$$Y_t=T_t+S_t+E_t \qquad (3.1)$$

$$\text{or } Y_t=T_t \cdot S_t \cdot E_t, \qquad (3.2)$$

where, Y_t=the original TS data, T_t=Trend component, S_t=Seasonal component and E_t=Error / Irregular component.

Equation (3.1) is called an additive seasonal model; this is appropriate for a time series in which the amplitude of the seasonal pattern is independent of the average level of the series, i.e. a time series displaying additive seasonality.

Equation (3.2) is called a multiplicative seasonal model; this is appropriate for a time series in which the amplitude of the seasonal pattern is proportional to the average level of the series, i.e. a time series displaying multiplicative seasonality. In other words, if the magnitude of a TS varies with the level of the series then one has to go for a multiplicative model, otherwise called an additive model. This decomposition may enable one to study the TS components separately or will allow analysts to de-trend or to do seasonal adjustments, if needed, for further analysis.

ARIMA Models

ARIMA is an abbreviation of AutoRegressive Integrated Moving Average introduced by Box and Jenkins (Box et.al., 1994). As such, some authors refer to this modelling approach as **Box and Jenkins model**. Box-Jenkins model is a stationary time series model. Time series that generated from zero-mean, finite variance, and uncorrelated variable are called ‘white noise’. Many useful models can be constructed from them.

Most physical processes exhibit inertia and do not change that quickly. This, combined with the sampling frequency, often makes consecutive observations correlate. Such correlation between consecutive observations is called autocorrelation. When the data are autocorrelated, most of the standard modelling methods, based on the assumption of independent observations may become misleading or sometimes even useless. We, therefore, need to consider alternative methods that take into account the serial dependence in the data. This can be fairly easily achieved by employing time series models such as the autoregressive integrated moving average (ARIMA) models.

ARIMA, which is often called method of Box-Jenkins time series, has appreciable accuracy for short-term forecasting, but less accuracy for long-term forecasting. Usually, it will tend to become flat for a sufficiently long period. ARIMA model ignores the independent variable completely, and uses past and present values of dependent variable to produce accurate short-term forecasting (Hendranata, 2003). ARIMA is suitable when the observation of time series is statistically related to the dependent. The purpose of this model is to determine good statistical relationships between the variables that are being predicted and the historical value of these variables, so that forecasting can be performed with the model (Hendranata, 2003).

The ARIMA modelling is essentially an exploratory data-oriented approach that has the flexibility of fitting an appropriate model which is adapted from the structure of the datum itself. The stochastic nature of the time series can be approximately modelled with the aid of autocorrelation function and partial autocorrelation function from which information such as trend, random variables, periodic components, cyclic patterns and serial correlation can be discovered. As a result, forecasts of the future values of the series with some degree of accuracy can be readily obtained (Ho and Xie, 1998).

Although ARIMA modelling is sophisticated in theory, but with the advent of computer technology, the iterative model building process and, hence, accurate forecast, can be aided and made simpler by the ease of many user-friendly statistical software packages such as SAS, SPSS, Statgraphics, Statistica and Minitab. An iterative three stage process, through model identification, parameter estimation and diagnostic check, is required to determine the adequacy of the proposed model (Ho and Xie, 1998).

ARIMA contains three components, namely: AutoRegressive (AR), Integrated (I) and Moving Average (MA) parts. The AR part describes the relationship between present and past observations. The MA part represents the autocorrelation structure of error. The I part represents the differencing level of the series to eliminate non-stationary (Hasmida, 2009). It is usually denoted by (p,d,q)(P,D,Q) where p denotes order of autoregressive component, d denotes order of differencing, q denotes order of moving average and (P,D,Q) denotes corresponding seasonal component.

AR(p) model expresses that the current value of time series as a linear combination of p previous values and a white noise term (random shock). Bell (1984) expresses the current value of time series of AR(p) model as:

$$Y_t = \phi_1 Y_{t-1} + \cdots + \phi_p Y_{t-p} + \varepsilon_t \qquad (3.3)$$

where ϕ_1 ,..., ϕ_p are AR(p) parameters, the ε_t is the random shock in normal distribution with zero mean and variance at time t, and p is the order of AR(p).
By introducing the backshift operator B, which defines (BY_t=Y_{t-1}), equation (3.3) can be written as:

$$\left(1 - \phi_1 B - \cdots - \phi_p B^p\right) Y_t = \varepsilon_t \qquad (3.4)$$

Or $\phi_p(B) Y_t = \varepsilon_t$ where $\phi_p(B) = \left(1 - \phi_1 B - \cdots - \phi_p B^p\right)$ (3.5)

MA(q) model expresses the current value of a time series as a linear combination of a current and q previous values of a white noise process. The (purely) moving average (MA) model is (Bell, 1984):

$$Y_t = \varepsilon_t - \theta_1 \varepsilon_{t-1} - \cdots - \theta_q \varepsilon_{t-q} \qquad (3.6)$$

Or $Y_t = \left(1 - \theta_1 B - \cdots - \theta_q B^q\right) \varepsilon_t$ or $Y_t = \theta_q(B) \varepsilon_t$ (3.7)

where q is the order of MA(q), and $\theta_q(B)$ are parameters of MA(q).

To increase flexibility when fitting actual time series, both autoregressive and moving average operators are combined to give the ARMA (p,q) model (Bell, 1984): coefficients are MA(q) model parameters.

$$Y_t = \phi_1 Y_{t-1} + \cdots + \phi_p Y_{t-p} + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \cdots - \theta_q \varepsilon_{t-q} \quad (3.8)$$

Which can be written as:

$$\left(1 - \phi_1 B - \cdots - \phi_p B^p\right) Y_t = \left(1 - \theta_1 B - \cdots - \theta_q B^q\right) \varepsilon_t \quad \text{or} \quad \phi_p(B) Y_t = \theta_q(B) \varepsilon_t \quad (3.9)$$

The mixed type of series which are explained both by its own lagged values and by lagged noise terms, is called Autoregressive Moving-Average models of order (p,q).

This systematic class of stationary time series models carries great importance and usefulness especially in real-life situations. If the process is stationary, a suitable ARMA model can be used to represent the data. If it is nonstationary, differencing is applied to make the model become stationary and this leads to ARIMA model (Akgun, 2003).

The first and most important condition of the series Y_t in equation (3.8) or (3.9) is that it has to be stationary. In practice Y_t may well be nonstationary, but has to be made stationary by first difference,

$$\nabla Y_t = Y_t - Y_{t-1} = (1-B)Y_t \qquad (3.10)$$

If equation (3.10) ie $(1\text{-B})Y_t$ is nonstationary, then there will be the need to take the second difference,

$$\nabla^2 Y_t = Y_t - 2Y_{t-1} + Y_{t-2} = (1-B)^2 Y_t \qquad (3.11)$$

In general, we may need to take the d^th^ difference $(1\text{-B})dY_t$ (although rarely is d larger than 2). Substituting $(1\text{-B})^dY_t$ for Y_t in (3.9) yields the ARIMA (p,d,q) model (Bell, 1984):

$$\phi_p(B)(1-B)^d Y_t = \theta_q(B)\varepsilon_t \qquad (3.12)$$

where d is the order of differencing.

When a time series exhibits potential seasonality indexed by s, using a multiplied seasonal ARIMA(p,d,q)(P,D,Q)^s model is advantageous. The seasonal time series is transformed into a stationary time series with non-periodic trend components. A multiplied seasonal ARIMA model can be expressed as (Lee and Ko, 2011):

$$\phi_p(B)\Phi_\mathrm{P}(\mathrm{B^s})\nabla^\mathrm{d}\nabla_s^D Y_t = \theta_q(B)\Theta_Q(B^s)\varepsilon_t \qquad (3.13)$$

where

$\phi_p(B) = 1 - \phi_1 B - \phi_2 B^2 \cdots - \phi_p B^p, \theta_q(B) = 1 - \theta_1 B - \theta_2 B^2 \cdots - \theta_q B^q$

$\Phi_\mathrm{P}(\mathrm{B^s}) = 1 - \Phi_1 B^s - \Phi_2 B^{2s} \cdots - \Phi_P B^{Ps},\ \Theta_Q(B^s) = 1 - \Theta_1 B^s - \Theta_2 B^{2s} \cdots - \Theta_Q B^{Qs}$

B is the backshift operator (i.e. $BY_t=Y_{t-1}$, $B^2Y_t=Y_{t-2}$, $B^{12}Y_t=Y_{t-12}$ and so on) s, is the seasonal lag, and ε_t is a sequence of independent normal error variables with mean zero and variance σ^2. $\nabla Y_t = Y_t - Y_{t-1} = (Y_t - BY_t) = (1-B)Y_t$, $\nabla^2 Y_t = \nabla(Y_t -$

$Y_{t-1}) = \nabla Y_t - \nabla Y_{t-1} = Y_t - Y_{t-1} - (Y_{t-1} - Y_{t-2}) = Y_t - 2Y_{t-1} + Y_{t-2} = (1 - 2B + B^2)Y_t = (1-B)^2 Y_t,\ \nabla_s Y_t = Y_t - Y_{t-s} = (1 - B^s)Y_t$

where D is the order of seasonal differencing, $\Phi_P(B^s)$ and $\Theta_Q(B^s)$ are the seasonal AR(p) and MA(q) operators respectively.

EXPONENTIAL SMOOTHING

Exponential smoothing is a procedure for continually revising a forecast in the light of more recent experience. Exponential Smoothing assigns exponentially decreasing weights as the observation gets older. In other words, recent observations are given relatively more weight in forecasting than the older observations.

Single Exponential Smoothing

This is also known as simple exponential smoothing. Simple smoothing is used for short-range forecasting, usually just one month into the future. The model assumes that the data fluctuates around a reasonably stable mean or level (no trend or consistent pattern of growth).
The specific formula for simple exponential smoothing is:

$$L_t = \alpha Y_t + (1 - \alpha)L_{t-1} \qquad (3.15)$$

where α is a smoothing constant between 0 and 1.

When applied recursively to each successive observation in the series, each new smoothed value (forecast) is computed as the weighted average of the current observation and the previous smoothed observation; the previous smoothed observation was computed in turn from the previous observed value and the smoothed value before the previous observation, and so on.

Thus, in effect, each smoothed value is the weighted average of the previous observations, where the weights decrease exponentially depending on the value of parameter (α). If it is equal to 1 (one) then the previous observations are ignored entirely; if it is equal to 0 (zero), then the current observation is ignored entirely, and the smoothed value consists entirely of the previous smoothed value (which in turn is computed from the smoothed observation before it, and so on; thus all smoothed values will be equal to the initial smoothed value L_0). In-between values will produce intermediate results.

Double Exponential Smoothing

This method is used when the data show a trend. Exponential smoothing with a trend works much like simple smoothing except that two components must be updated each period - level and trend. The level is a smoothed estimate of the value of the data at the end of each period. The trend is a smoothed estimate of average growth at the end of each period. The specific formulas for double exponential smoothing are:

$$L_t = \alpha Y_t + (1-\alpha)(L_{t-1} + T_{t-1}) \quad 0<\alpha<1 \qquad (3.16)$$

$$T_t = \gamma(L_t - L_{t-1}) + (1-\gamma)T_{t-1} \quad 0<\gamma<1 \qquad (3.17)$$

Note that the current value of the series is used to calculate its smoothed value replacement in double exponential smoothing.

Triple Exponential Smoothing

This method is used when the data show level, trend and seasonality. To handle seasonality, we have to add a third parameter. We now introduce a third equation to take care of seasonality. The resulting set of equations is called the ”Holt-Winters” (HW) method, after the names of the inventors.

Estimate of the level $L_t = \alpha(Y_t/S_{t-s}) + (1-\alpha)(L_{t-1} + T_{t-1})$ (3.18)

Estimate of the growth rate (or trend) $T_t = \gamma(L_t - L_{t-1}) + (1-\gamma)T_{t-1}$ (3.19)

Estimate of the seasonal factor $S_t = \delta(Y_t/L_t) + (1-\delta)S_{t-s}$ (3.20)

where α, γ, and δ are smoothing constants between 0 and 1, s = number of seasons in a year (s = 12 for monthly data, and s = 4 for quarterly data).

GOODNESS-OF-FIT MEASURES

This subsection provides definitions of the goodness-of-fit measures used in time series modelling.

1. *Stationary R-squared*. A measure that compares the stationary part of the model to a simple mean model. This measure is preferable to ordinary R-squared when there is a trend or seasonal pattern. Stationary R-squared can be negative with a range of negative infinity to 1. Negative values mean that the model under consideration is worse than the baseline model. Positive values mean that the model under consideration is better than the baseline model.

2. *R-squared*. An estimate of the proportion of the total variation in the series that is explained by the model. This measure is most useful when the series is stationary. R-squared can be negative with a range of negative infinity to 1. Negative values mean that the model under consideration is worse than the baseline model. Positive values mean that the model under consideration is better than the baseline model.
3. *RMSE*. Root Mean Square Error. The square root of mean square error. A measure of how much a dependent series varies from its model-predicted level, expressed in the same units as the dependent series.
4. *MAPE*. Mean Absolute Percentage Error. A measure of how much a dependent series varies from its model-predicted level. It is independent of the units used and can, therefore, be used to compare series with different units.
5. *MAE*. Mean absolute error. Measures how much the series varies from its model-predicted level. MAE is reported in the original series units.
6. *MaxAPE*. Maximum Absolute Percentage Error. The largest forecast error, expressed as a percentage. This measure is useful for imagining a worst-case scenario for your forecasts.
7. *MaxAE*. Maximum Absolute Error. The largest forecast error, expressed in the same units as the dependent series. Like MaxAPE, it is useful for imagining the worst-case scenario for your forecasts. Maximum absolute error and maximum absolute percentage error may occur at different series points - for example, when the absolute error for a large series value is slightly larger than the absolute error for a small series value. In that case, the maximum absolute error will occur at the larger series value and the maximum absolute percentage error will occur at the smaller series value.
8. *Normalized BIC*. Normalized Bayesian Information Criterion. A general measure of the overall fit of a model that attempts to account for model complexity. It is a score based upon the mean square error and includes a penalty for the number of parameters in the model and the length of the series. The penalty removes the advantage of models with more parameters, making the statistic easy to compare across different models for the same series.
9. *Significance Level (p-value)*: There is always a probabilistic component involved in the accept–reject decision in testing hypothesis. The criterion that is used for accepting or rejecting a null hypothesis is called significance level or p-value. The p-value represents the probability of concluding (incorrectly) that there is a difference in your samples when no true difference exists. In other words, a p-value of 0.05 means there is only a 5% chance that you would be wrong in concluding that the populations are different or 95% confident of making a right

decision. For atmospheric sciences research, a p-value of 0.05 or 0.10 is generally taken as standard.

EXAMPLE: ANALYSIS OF METROLOGICAL DATA

Daily data used in this presentation were obtained from Centre for Atmospheric Research (CAR), sited at Kogi State University Campus, Anyigba, Nigeria. The station has in its data base, four the meteorological parameters of solar radiation, relative humidity, temperature and wind speed, daily data spanning for three years (2010, 2011 and 2012). The data, which was recorded at five minutes intervals, were averaged monthly for sunshine hours between 07.00 and 18.00 hours local time, using Microsoft Excel spread sheet.

Methodology

An important preliminary step in any data analysis is to consider the possibility of (non-linear) data transformation. It is often the case that the scale in which the data naturally arrives to the data analyst is not necessarily the best scale to analyse them. The primary goal of the transformation is to identify a scale where the residuals, after fitting a model, will have homogeneous variability and be independent of the level of the time series. The most common transformation used is the logarithmic transformation.

In this case, the Expert Modeler of SPSS 16.0 software was used. The Expert Modeler only selects the candidate predictors to find the best model of those predictors that have a statistically significant relationship with the dependent series. It shows how predictors are useful (in terms of how each predictor is significant) and the model developed can be used for making forecast with the predictors. The modeler will give either exponential smoothing or ARIMA models. It shows whether the model is additive or multiplicative and also whether there is/are transformations.
Forecasts are made of these parameters for the year 2013.

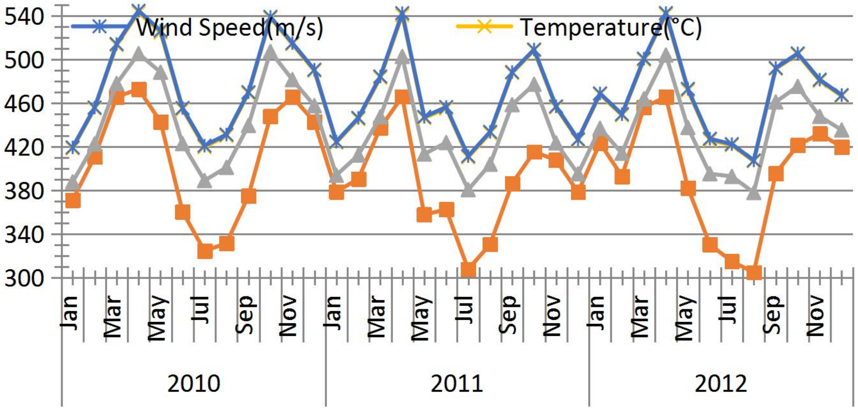


Figure 3.1: *The plots of the meteorological data for the years 2010 to 2012 for the area under study*

Figure 3.1 shows the behaviours of the parameters. It can be observed that the parameters are almost proportional to each other, because they almost follow the same cyclic pattern. It can be observed that solar radiation, wind speed, temperature, and RH are correlated. Wind speed and temperature follow the same trend in all the months of the three years period. RH and, in turn, Solar radiation are higher in the months between March and June, dipping between July and August and higher again in September to October and low in November and December. This trend is normal because solar radiation in the months of March to June are expected to be high as a result of clear sky and lower dust loadings due to the wind movement from land to sea. The lower values are experienced during the rainy season, when the skies are cloudy and wind movement is from sea to land. This will result in the absorption of incoming solar radiation by water droplets, leading to its extinction before reaching recording instruments.

Results and Discussions

Solar Radiation

Table 3.1: *Model Summary of Model Parameters for Solar Radiation*

Solar_Rad	Model	Simple Seasonal	
Stationary R^2	R^2	Sig.	
0.81793	0.86173	0.12010	
Exponential Smoothing Model Parameters			
No Transformation		Estimate	Sig.
	Alpha (Level)	0.099998	0.202155
	Delta (Season)	0.000091	0.999698

From Table 3.1, the model obtained was simple seasonal. This model is appropriate for series with no trend and a seasonal effect that is constant over time. Its smoothing parameters are level and season. This shows that solar radiation did not increase over the years and its seasonal effect is constant for these years. From the values of R^2 and stationary R^2, it can be said that the model is good, but by observing the value of significance, it shows that the model is not significant. Also, from the values of the significant of the model parameters, it can be seen that the parameters are not significant, most especially, the seasonal parameter.

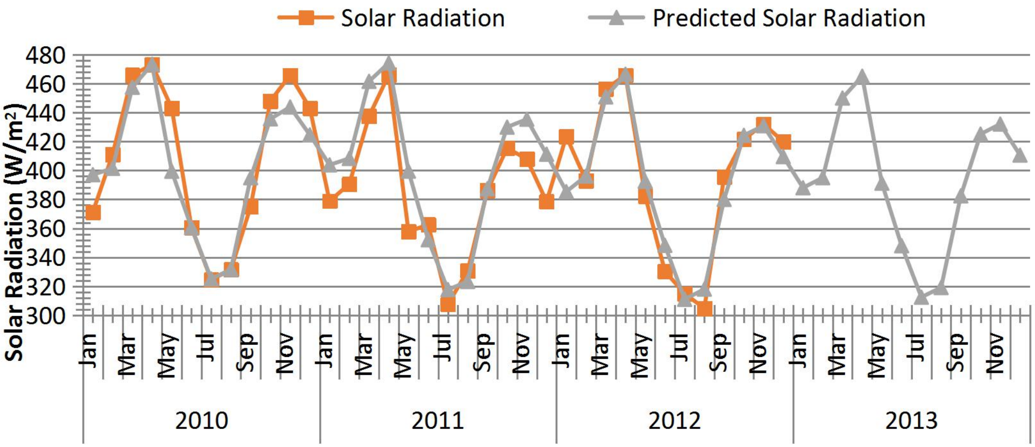


Figure 3.2: *The plots of measured solar radiation and estimated solar radiation using time series analysis*

Figure 3.2 shows the plots of solar radiation for the measured and the estimated from our modelling. The plots show that the model can fairly estimate the parameter very well.

Relative Humidity

Table 3.2: Model Summary of Model Parameters for relative humidity

RH	**Model**	**Simple Seasonal**	
Stationary R^2	R^2	Sig.	
0.78795	0.98466	0.59696	
Exponential Smoothing Model Parameters			
No Transformation		Estimate	Sig.
	Alpha (Level)	0.199978738	0.069059
	Delta (Season)	9.95326E-06	0.999904

From Table 3.2, the model obtained was simple seasonal. This model is appropriate for series with no trend and a seasonal effect that is constant over time. Its smoothing parameters are level and season. This shows that RH did not increase over the years and its seasonal effect is constant for these years. From the values of R^2 and stationary R^2, it can be said that the model is good, but by observing the value of significance, it is clear that the model is not significant. Also, from the values of the significant of the model parameters, it can be seen that the parameters are not significant, most especially, the seasonal parameter.

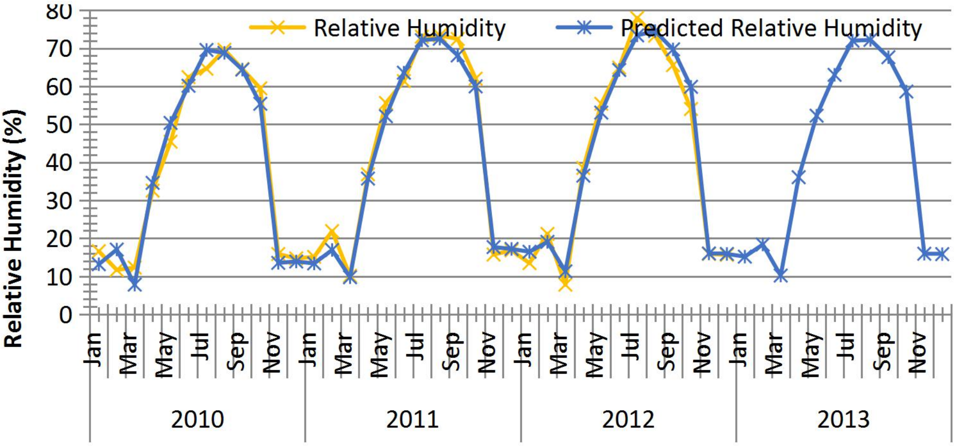


Figure 3.3: *The plots of measured Relative Humidity and Estimated Relative Humidity using time series analysis*

Figure 3.3 shows the plots of RH for the measured and the estimated from our modelling. The plots show that the model can estimate the parameter very well.

Temperature

Table 3.3: *Model summary of model parameters for temperature*

Temp	**Model_3**	**Simple Seasonal**	
Stationary R^2	R^2	Sig.	
0.83898	0.93209	0.07129	
Exponential Smoothing Model Parameters			
No Transformation		Estimate	Sig.
	Alpha (Level)	0.09979	0.349563
	Delta (Season)	5.67E-05	0.999747

From Table 3.3, the model obtained was simple seasonal. This model is appropriate for series with no trend and a seasonal effect that is constant over time. Its smoothing parameters are level and season. This shows that temperature did not increase over the years and its seasonal effect is constant for these years. From the values of R^2 and stationary R^2, it can be said that the model is good, but by observing the value of significance, it shows that the model is not significant. Also, from the values of the significant of the model parameters, it can be seen that the parameters are not significant, most especially, the seasonal parameter.

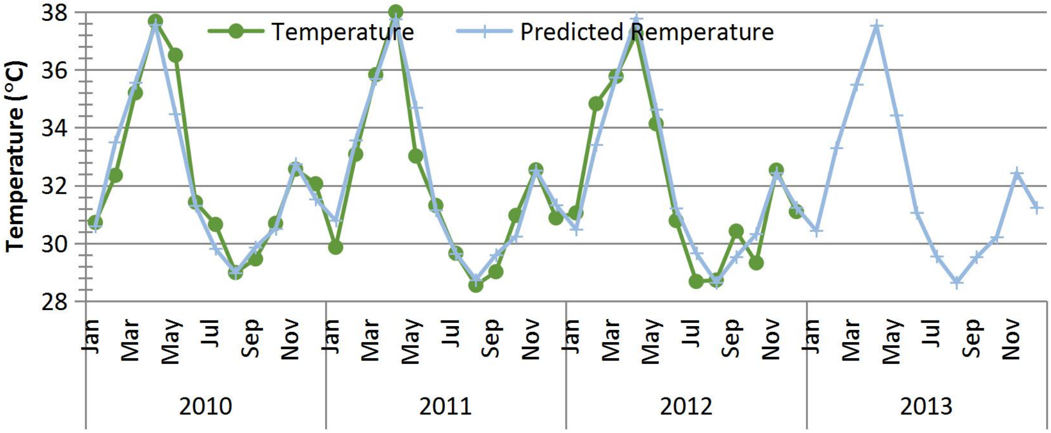


Figure 3.4: *The plots of measured Temperature and Estimated measured Temperature using time series analysis*

Figure 3.4 shows the plots of temperature for the measured and the estimated from our modelling. The plots show that the model can estimate the parameter very well.

Wind Speed

Table 3.4: *Model summary of model parameters for wind speed.*

WS	**Model_4**	**Winters' Additive**	
Stationary R^2	R^2	Sig.	
0.83230	0.73496	0.01348	
Exponential Smoothing Model Parameters			
No Transformation		Estimate	Sig.
	Alpha (Level)	0.088699597	0.570853896
	Gamma (Trend)	3.02474E-07	0.999992013
	Delta (Season)	4.26079E-05	0.999860413

From Table 3.4, the model obtained was Winter's additive. This model is appropriate for series with a linear trend and a seasonal effect that does not depend on the level of the series. Its smoothing parameters are level, trend, and season. This shows that wind speed has a linear trend and a seasonal effect that does not depend on the level of the level of the speed (that is it does not depend on the initial speed). From the values of R^2 and stationary R^2, it can be said that the model is good, and also by observing the value of significance, it shows that the model is very significant. Also, from the values of the significant of the model parameters, it can be seen that the parameters are not significant, most especially, the trend and seasonal parameters.

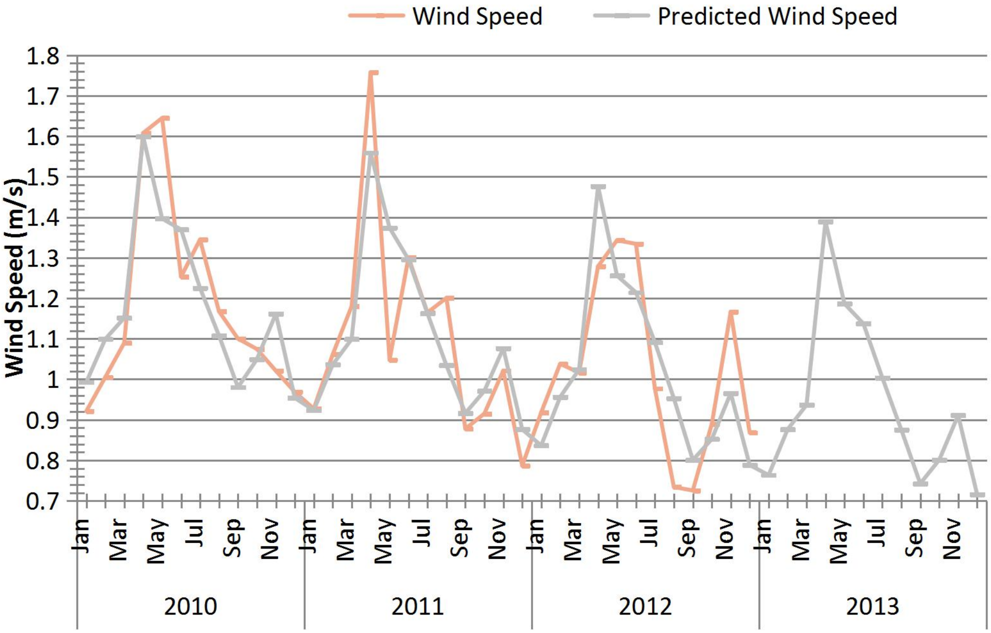


Figure 3.5: *The plots of measured Wind Speed and Estimated measured Wind Speed using time series analysis*

Figure 3.5 shows the plots of wind speed for the measured and the estimated from our modelling. The plots show that the model can only estimate the parameter poorly.

Conclusion

From the analysis of the results, it can be said that solar radiation, relative humidity and temperature have simple seasonal relation. That is they do not increase throughout the three years, and their seasonal effects are constants. The wind speed has winters additive model. That is, it has linear trend and the seasonal effect does not depend on the level (initial speed). Also, considering the values of the stationary R^2, it can be said that the models are good, since each one can estimate more than 75% of the measured parameters. The American Statistical Association (ASA) has released a "Statement on Statistical Significance and P-Values" with six principles underlying the proper use and interpretation of the p-value. It was finally concluded that the p-value should never be intended to be a substitute for scientific reasoning and that well-reasoned statistical arguments should contain much more than the value of a single number (ASA News, 2016).

EMPIRICAL ORTHONAL FUNCTIONS

Introduction

Climate can be defined, in mathematical terms, as the aggregation of all long-term statistical properties of the atmospheric state or, as defined by Lorenz (1970), Climate is regarded as the aggregation of (random) daily weather. It is, therefore, the long-term statistics of weather. Climate variations are also the result of exceedingly complex non-linear interactions between very many degrees of freedom or modes. Both weather and climate are characterized by non-linearity and high dimensionality. Consequently, a challenging task is to find ways to reduce the dimensionality of the system to a few modes if possible. A further, yet challenging, task is to link these modes to the dynamics/physics of the system.

Empirical orthogonal function (EOF) analysis, also called principal component analysis (PCA) (Fukuoka, 1951; Lorenz, 1956), is among the most widely and extensively used methods in dimensionality reduction and patterns extraction in atmospheric science. Both names are commonly used, and refer to the same set of procedures. Typically, the EOFs are found by computing the eigenvalues and eigenvectors of a spatially weighted anomaly covariance matrix of a field. The derived eigenvalues provide a measure of the percent variance explained by each mode.

The atmospheric data are usually converted into a two-dimensional matrix. The easiest example to imagine is a data set that consists of observations of several variables at one instant of time, but includes many realizations of these variable values taken at different times. One can imagine several possible generic types of data matrices.

i) A space-time array: Measurements of a single variable at M locations taken at N different times, where M and N are integers.

ii) A parameter-time array: Measurements of M variables (e.g. temperature, pressure, relative humidity, rainfall, . . .) taken at one location at N times.

iii) A parameter-space array: Measurements of M variables taken at N different locations at a single time.

iv) EOF analysis could just as well be applied to concentrations of M different chemical compounds from N different experiments.

One can still imagine other possibilities.

This presentation is intended to provide a basic introduction to what has become a very large subject. The theories presented here are the main rudimentary aspects that are needed in understanding the EOF. The SPSS was used in the analysis, and here demonstrations were done on how to analyse the outputs.

THEORY

The data to be used for the EOF analysis are transformed into a two-dimensional data matrix by the software, X (say) as follows:

$$X = \begin{array}{cc} & N \\ M & \left[\qquad \right] \end{array} = X_{i,j}\ where\ i = 1{,}M; j = 1{,}N \qquad (4.1)$$

where M and N are the dimensions of the data matrix enclosed by the square brackets, and the subscript notation $X_{i,j}$ to indicate the same matrix. The transpose of the matrix are obtained by reversing the order of the indices to make it an NxM matrix.

$$X^T = \begin{array}{cc} & M \\ N & \left[\qquad \right] \end{array} = X_{j,i}\ where\ i = 1{,}M; j = 1{,}N \qquad (4.2)$$

In multiplying a matrix times itself we generally need to transpose it once to form an inner product, which results in two possible “dispersions” matrices, in this case are called correlation/covariance matrices. The first can be obtained as:

$$C = XX^T = \begin{array}{ccc} & N & M \\ M & \left[\qquad \right] & \left[\qquad \right] \end{array} N = X_{i,j}X_{j,i} = X_{i,i} = \begin{array}{cc} M & \\ \left[\qquad \right] & M \end{array} \qquad (4.3)$$

The other dispersion matrix in which the roles of the structure and sampling variables are reversed.

$$C = X^TX = \begin{array}{ccc} & M & N \\ N & \left[\qquad \right] & \left[\qquad \right] \end{array} M = X_{j,i}X_{i,j} = X_{j,j} = \begin{array}{cc} N & \\ \left[\qquad \right] & N \end{array} \qquad (4.4)$$

In this projection of a matrix onto itself, one of the dimensions gets removed and we are left with a measure of the dispersion of the structure with itself across the removed dimension (or the sampling dimension). If the sampling dimension is time, then the resulting dispersion matrix is the matrix of the covariance of the spatial locations with each other, as determined by their variations in time. These dispersion matrices are in fact covariance/correlation matrices. In the second case, the covariance at different times is obtained by projecting on the sample of different spatial points. Either of these dispersion matrices may be scientifically meaningful, depending on the problem under consideration. These matrices generated can be either positive definite or positive semi-definite covariance/correlation matrices and are usually symmetrical.

EOF (or PCA) analysis consists of an eigenvalue analysis of any one of these dispersion matrices. Any symmetric matrix C can be decomposed through a diagonalization, or eigen analysis using the following:

$$Ce_i = \lambda_i e_i \qquad (4.6)$$

$$CE = E\Lambda \qquad (4.7)$$

where E is the matrix with the eigenvectors e_i as its columns, and Λ is the matrix with the eigenvalues λ_i, along its diagonal and zeros elsewhere. The eigenfucntions of the Hermtian matrix C form an orthonormal basis to represent C and hence called empirical functions or empirical modes. The eigenvalue λ_i is a measure of the percentage variability represented by ith EOF.

The set of eigenvectors, e_i, and associated eigenvalues, λ_i, represent a coordinate transformation into a coordinate space where the matrix C becomes diagonal. Because the covariance/correlation matrix is diagonal in this new coordinate space, the variations in these new directions are uncorrelated with each other, at least for the sample that has been used to construct the original covariance/correlation matrix. The eigenvectors define directions in the initial coordinate space along which the maximum possible variance can be explained, and in which variance in one direction is orthogonal to the variance explained by other directions defined by the other eigenvectors. The eigenvalues indicate how much variance is explained by each eigenvector. If you arrange the eigenvector/ eigenvalue pairs with the biggest eigenvalues first, then you may be able to explain a large amount of the variance in the original data set with relative few coordinate directions, or characteristic structures in the original structure space.

The tendency of the empirical modes to extract poorly representative commonality among subdomains of large datasets can be remedied by grouping the variance through a rotation procedure. A variety of such procedures are available (Richman, 1986); however, the rotation technique most commonly used to group the variability in geophysical applications is the varimax orthogonal rotation.

Most of the rationale for rotating factors comes from Thurstone (1947) and Cattell (1978) who defended its use because this procedure simplifies the factor structure and therefore makes its interpretation easier and more reliable (i.e., easier to replicate with different data samples).

A rotation is specified by a rotation matrix denoted R, where the rows stand for the original factors and the columns for the new (rotated) factors. At the intersection of row m and column n we have the cosine of the angle between the original axis and the new one: $r_{m,n}=\cos\theta_{m,n}$. For example the rotation illustrated in Figure 4.1 will be characterized by the following matrix:

$$\mathrm{R} = \begin{bmatrix} \cos\theta_{1,1} & \cos\theta_{1,2} \\ \cos\theta_{2,1} & \cos\theta_{2,2} \end{bmatrix} = \begin{bmatrix} \cos\theta_{1,1} & -\sin\theta_{1,1} \\ \sin\theta_{1,1} & \cos\theta_{1,1} \end{bmatrix} \qquad (4.8)$$

with a value of $\theta_{1,1}$=15 degrees. A rotation matrix has the important property of being orthonormal because it corresponds to a matrix of direction cosines and therefore $R^TR=I$.

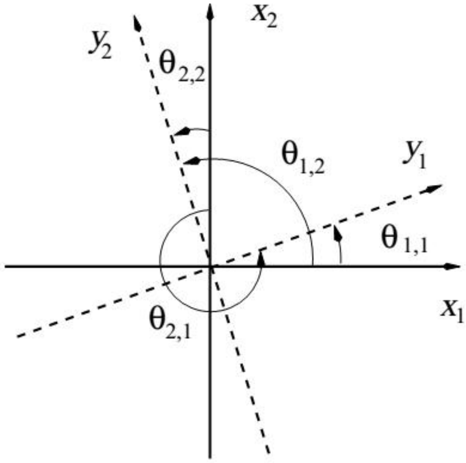


Figure 4.1: An orthogonal rotation in 2 dimensions. The angle of rotation between an old axis m and a new axis n is denoted by $\theta_{m,n}$.

In oblique rotations the new axes are free to take any position in the factor space, but the degree of correlation allowed among factors is, in general, small because two highly correlated factors are better interpreted as only one factor. Oblique rotations, therefore, relax the orthogonality constraint in order to gain simplicity in the interpretation. They were strongly recommended by Thurstone, but are used more rarely than their orthogonal counterparts.

In general, a rotation is a linear transformation of the modes that attempts to find a new location for the coordinate axis, such that projections of the variable onto those axes simplify the spatial or temporal structure of the modes. A detailed discussion of the advantages and disadvantages of rotated empirical modes is given by Richman (1986); Jolliffe, (1987); Richman, (1987). In most applications, the rotation is used to simplify the spatial structure by isolating regions with similar temporal variability (e.g. Horel, 1981; Barnston and Livezey, 1987; Kawamura, 1994; Mestas-Nunez and Enfield, 1999). The resulting rotated space patterns are generally more robust (i.e. less sensitive to sampling errors) than their unrotated counterparts (Cheng *et. al.*, 1995). Alternatively, the rotation can also be used to simplify the temporal structure by isolating time periods with similar space patterns (e.g. Fernandez, 1995).

Interpretation is more straightforward if each variable is highly loaded on at most one factor, and if all factor loadings (also known as correlation coefficients) are either large or positive or near zero, with few intermediate values (see Everitt and Dunn, 2001). The SPSS provides several methods of rotation that try to achieve these goals, some of which produce orthogonal factors (varimax, quartimax, and equamax) and others that lead to an oblique solution (direct oblimin and promax).
The following are the six possible options for the rotations:

(i) The first one no rotation.

(ii) **Varimax Method -** An orthogonal rotation method that minimizes the number of variables that have high loadings on each factor. This method simplifies the interpretation of the factors.

(iii) **Quartimax Method** - A rotation method that minimizes the number of factors needed to explain each variable. This method simplifies the interpretation of the observed variables.

(iv) **Equamax Method** - A rotation method that is a combination of the varimax method, which simplifies the factors, and the quartimax method, which simplifies the variables. The number of variables that load highly on a factor and the number of factors needed to explain a variable are minimized.

(v) **Direct Oblimin Method** - A method for oblique (nonorthogonal) rotation. When delta equals 0 (the default), solutions are most oblique. As delta becomes more negative, the factors become less oblique. To override the default delta of 0, enter a number less than or equal to 0.8.

(vi) **Promax Rotation** - An oblique rotation, which allows factors to be correlated. This rotation can be calculated more quickly than a direct oblimin rotation, so it is useful for large datasets.

For orthogonal rotations, the rotated component matrix and component transformation matrix are displayed. For oblique rotations, the pattern, structure and component correlation matrices are displayed. In addition they all have Component Score Coefficient Matrix (Eigen Vectors) displayed.

Example

In this following example, the non-rotated EOF and the five rotated EOFs are analysed. The data used here are from the meteorological data used in the time series (Section 3).

Results and Discussions

Here are the results of the analysis using empirical orthogonal functions:

Table 4.1: *The correlation matrix*

		Solar_Rad	**RH**	**Temp**	**WS**
Correlation	Solar_Rad	1	-0.6428	0.7129	0.2002
	RH	-0.6428	1	-0.5044	0.1084
	Temp	0.7129	-0.5044	1	0.6027
	WS	0.2002	0.1084	0.6027	1

Table 4.1 shows the correlation matrix. This is typically used to do an eyeball test and to get a feeling for which variable is strongly associated with which variable. The correlation matrix is closely related to multiple regression and explained variance. The matrix shows the fraction of the variance of each variable that can be explained by all of the other variables. Since this is relatively large, it suggests that the variables are closely related and that the data set is, therefore, a good candidate for factor analysis. The off diagonal terms are the fractions of the variance of each variable that can only be explained by the variable indicated. It shows large correlations between solar radiation with RH and temperature, between RH and temperature, temperature and wind speed and low correlation between wind speed with solar radiation and RH.

Table 4.2: *The communalities of the parameters*

	Initial	Extraction
Solar_Rad	1	0.83090131
RH	1	0.87113246
Temp	1	0.91835973
WS	1	0.95109195

From Table 4.2, it is clear that the communalities are found from the factor solution by the sum of the squared loadings. The Table shows that 83.1% of the solar radiation, 87.1% of the RH, 91.8% of temperature and 95.1% of wind speed are all accounted by the components extracted.

Table 4.3: The KMO and Bartlett's Test

KMO and Bartlett's Test		
Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		0.5358
Bartlett's Test of Sphericity	Approx. Chi-Square	71.5520
	Df	6
	Sig.	1.96E-13

Table 4.3 is the KMO and Bartlett test that is used for the test of sphericity. The KMO criterion can have values between [0,1] where the usual interpretation is that 0.8 indicates a good adequacy to use the data in a factor analysis. If the KMO criterion is less than 0.5, this implies that no meaningful information can be obtained. From the value of the KMO it is evident that some information can be obtained. The value of the significance parameter shows that the data are statistically significant.

i) No Rotation

Table 4.4: *The total variance explained table for no rotation*

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	2.3774	59.4342	59.4342	2.3774	59.4342	59.4342
2	1.1941	29.8529	89.2871	1.1941	29.8529	89.2871
3	0.2921	7.3021	96.5892			
4	0.1364	3.4108	100			

Table 4.4 summarizes the total variance explained by the PCA solution and gives an indication about the number of useful factors. This Table has three parts. The first part shows the number of components extracted, which is usually the same as the number of the variables or parameters. There are a total of four (4) components, which is the same as the number of variables entered into the PCA. However, please note that these components are not the same as the variables. The first column under **Initial Eigenvalues** gives the eigenvalues for all the possible factors in a decreasing order. This is followed by the variance as a percentage of all the variance and cumulative variance.

The third part, titled: **Extraction Sums of Squared Loadings**, gives information for factors with eigenvalues greater than 1. The word "extraction" here refers to the fact that these values are calculated after factor extraction. SPSS extracts all factors that have an eigenvalue greater than 1. In our own case, the analysis extracted two factors. This shows how much of the total variance of the observed variables is explained by each of the principal components. The first principal component (scaled eigenvector), by definition the one that explains the largest part of the total variance, has a variance (eigenvalue) of 2.38; this amounts to 59% of the total variance. The second principal component has a variance of 1.19 and accounts for 29.9% of the variance. The "Cumulative %" column of the Table tells us how much of the total variance can be accounted for by the first two(2) components put together. For example, the first two principal components account for 89.3% of the total variance.

Notice that the third and fourth eigenvalues are small. Pretty clearly there are only two significant eigenvectors here. The rule of thumb is that the model should explain more than 70% of the variance. In our data the model explains 89.3%.

Table 4.5: *Component Matrix for no rotation*

	Component	
	1	2
Temp	0.9274	0.2413
Solar_Rad	0.8827	-0.2276
RH	-0.7139	0.6012
WS	0.4779	0.8501

The correlations for the first two unrotated components loadings are shown in Table 4.5. These values represent how the variables are weighted for each component and the correlation between the variables and the components. For each of the variables, we get a loading in each of the columns representing components. The variables are listed in the decreasing order of factor loadings as we requested the same in the Options window.

The coefficients in this Table specify the linear function of the observed variables that define each component. The SPSS presents the coefficients scaled so that when the principal component analysis is based on the correlation matrix, they give the correlations between the observed variables and the principal components. These coefficients are often used to interpret the principal components and, possibly, give them names.

It can be observed that component 1 is mostly dominated by temperature, solar radiation, RH and moderate wind speed. This can be characterized as typical dry season. The second component is dominated by RH and wind speed, and this is the typical nature of rainy season.

In this Table 4.5, it can also be seen that the variables which have high correlation on component 1 have low correlation on component 2. Likewise, those with high correlation on component 2 have low correlation on component1.

Table 4.6: *Component Score Coefficient Matrix (Eigen Vectors) for no rotation*

	Component	
	1	2
Solar_Rad	0.3713	-0.1906
RH	-0.3003	0.5035
Temp	0.3901	0.2021
WS	0.2010	0.7119

Table 4.6 shows the eigenvectors of the two eigenvalues displayed. These values represent the coefficient of the basis vectors. These are the principal components.
Table 4.7: Reproduced Correlations after the analysis *for no rotation*

		Solar_Rad	**RH**	**Temp**	**WS**
Reproduced Correlation	Solar_Rad	0.8309	-0.7670	0.7637	0.2284
	RH	-0.7670	0.8711	-0.5171	0.1699
	Temp	0.7637	-0.5171	0.9184	0.6484
	WS	0.2284	0.1699	0.6484	0.9511

From Table 4.7 The reproduced correlation matrix is related to multiple regression and explained variance. The matrix shows the correlation coefficients of the variance of each variable that can be explained by all of the other variables. Since the coefficients are relatively large, it suggests that the variables are closely related and that the data set is, therefore, a good candidate for factor analysis. The off diagonal terms are the coefficients of the variance of each variable that can only be explained by the variable indicated.

ii) *Orthogonal Rotations*

Factor extraction is usually followed by rotation in order to maximize large correlations coefficients and minimize small correlations coefficients. Rotation usually increases simple structure and interpretability. The most commonly used is the Varimax variance maximizing procedure which maximizes the correlations coefficients.

(a) **Verimax**: this is focusing on the columns. It tends to produce multiple group components, maintaining orthogonality often results in increased multivocality (loadings of variables on "primary factors" is decreased a bit and loadings on "secondary factors" is raised a bit).

(b)

Table 4.8: *The Total Variance Explained for Verimax*

Component	Initial Eigenvalues			Rotation Sums of Squared Loadings		
	Total	% Var.	Cum. %	Total	% Var	Cum. %
1	2.3774	59.4342	59.4342	2.0529	51.3237	51.3237
2	1.1941	29.8529	89.2871	1.5185	37.9634	89.2871
3	0.2921	7.3021	96.5892			
4	0.1364	3.4108	100			

Table 4.8 summarizes the total variance explained by the PCA solution and gives an indication about the number of useful factors as explained previously.

The last part titled: **Rotated Sums of Squared Loadings** gives the information for extracted factors after rotation. It can be noted that after rotation, only the relative value of eigenvalues has changed, the cumulative percentage remains the same. The first principal component (scaled eigenvector), which explains the largest part of the total variance, has a variance (eigenvalue) of 2.05; this amounts to 51.32% of the total variance. The second principal component has a variance 1.52 and accounts for a further 37.96% of the variance. The first two principal components account for 89.29% of the total variance.

Table 4.9: The Rotated Component Matrix for Verimax

	Component	
	1	2
RH	-0.9230	0.1384
Solar_Rad	0.8712	0.2683
WS	-0.0379	0.9745
Temp	0.6638	0.6912

From Table 4.9 it is evident that if we perform an orthogonal rotation we obtain the correlation structure in which we find that the first component has strong correlations with the three physical variables (Solar radiation, temperature and RH), and a weak correlation with wind speed, and the second component has strong correlations with wind speed and temperature.

The rotated components of the varimax orthogonal rotation represent both how the variables are weighted for each factor and the correlation between the variables, and the components. A varimax rotation attempts to maximize the squared loadings of the columns. Based on these components loadings, it can be said that the first component represents dry season, and the second component represents rainy season. The complex loadings of temperature implies that the place had high temperature throughout the years.

Table 4.10 The Component Transformation Matrix for Verimax

Component	1	2
1	0.851953447	0.52361754
2	-0.523617536	0.85195345

Table 4.10 shows the component transformation matrix. This is used to see whether the rotation technique is suitable or not. Usually, a suitable rotation technique will result in a nearly symmetrical off-diagonal, which is not true in this case.

Table 4.11: The Component Score Coefficient Matrix (Eigen Vectors) for Verimax

	Component	
	1	2
Solar_Rad	0.4161	0.0320
RH	-0.5195	0.2717
Temp	0.2265	0.3764
WS	-0.2015	0.7118

Table 4.11 shows the eigenvectors of the two eigenvalues displayed in Table 4.8. The values represent the coefficients of the basis vectors. The first component shows that solar radiation and temperature have positive values, while RH and wind speed have negative values. From the second component, it can be seen that all have positive coefficients, but wind speed, temperature and RH, have higher coefficients while solar radiation has very small value.

b) **Quartimax**: this is focusing on the rows and it tends to produce a general factor and additional smaller multiple group factors

Table 4.12: Total Variance Explained for Quartimax

Component	Initial Eigenvalues			Rotation Sums of Squared Loadings		
	Total	% Var	Cum %	Total	% Var	Cum %
1	2.3774	59.434	59.434	2.143	53.567	53.567
2	1.1941	29.853	89.287	1.429	35.721	89.287
3	0.2921	7.302	96.589			
4	0.1364	3.411	100			

As reflected in Table 4.12, the last part titled: "Rotated Sums of Squared Loadings" gives the information for extracted components after rotation. Note that after rotation, only the relative value of eigenvalues has changed, the cumulative percentage remains the same.

The first principal component (scaled eigenvector), which explains the largest part of the total variance, has a variance (eigenvalue) of 2.14; this amounts to 53.57% of the total variance. The second principal component has a variance 1.43 and accounts for a further 35.72% of the variance. The first two principal components account for 89.29% of the total variance.

Table 4.13: Rotated Component Matrix for Quartimax

	Component	
	1	2
RH	-0.9070	0.2203
Solar_Rad	0.8917	0.1894
Temp	0.7229	0.6291
WS	0.0493	0.9740

From Table 4.13, after performing the orthogonal rotation, we obtain the correlation structure in which we find that the first component has strong correlations with the three physical variables (Solar radiation, temperature and RH), and a weak correlation

with wind speed, and the second component has strong correlations with wind speed and temperature.
The rotated component for the Quartimax orthogonal rotation represents both how the variables are weighted for each factor and the correlation between the variables, and the components. Based on these components loadings, it can be said that the first component represents dry season, and the second component represents rainy season. Because of the complex loadings of temperature, this implies that the place has high temperature throughout the years.

Table 4.14 Component Transformation Matrix for Quartimax

Component	1	2
1	0.8953	0.4454
2	-0.4454	0.8953

Table 4.14 shows the component transformation matrix. This is used to see whether the rotation technique is suitable or not. Usually, a suitable rotation technique will result in a nearly symmetrical off-diagonal which is not true in the case.

c) **Equamax**: this is a compromise between Verimax and Quartimax.

Table 4.15: Total Variance Explained for Equamax

Component	Initial Eigenvalues			Rotation Sums of Squared Loadings		
	Total	% Var	Cum %	Total	% of Var	Cum %
1	2.377	59.434	59.434	2.053	51.324	51.324
2	1.194	29.853	89.287	1.519	37.963	89.287
3	0.292	7.302	96.589			
4	0.1364	3.410	100			

As it is depicted in Table 4.15, the last part titled: "Rotated Sums of Squared Loadings" gives the information for extracted components after rotation.
The first principal component (scaled eigenvector) which explains the largest part of the total variance has a variance (eigenvalue) of 2.05; this amounts to 51.32% of the total variance. The second principal component has a variance 1.52 and accounts for 37.96% of the variance. The first two principal components account for 89.29% of the total variance.

Note that after rotation it is only the relative value of eigenvalues that has changed the cumulative percentage remains the same.

Table 4.16 :Rotated Component Matrix for Equamax

	Component	
	1	2
RH	-0.9230	0.1384
Solar_Rad	0.8712	0.2683
WS	-0.0379	0.9745
Temp	0.6638	0.6912

In Table 4.16, it is clear that after performing the orthogonal rotation, we obtain the correlation structure, in which we find that the first component has strong correlations with the three physical variables (Solar radiation, temperature and RH), and a weak correlation with wind speed; the second component has strong correlations with wind speed and temperature.

The rotated component for the Equamax orthogonal rotation represents both how the variables are weighted for each factor and the correlation between the variables, and the components. Based on these components loadings, it can be said that the first component represents dry season, and the second component represents rainy season. Because of the complex loadings of temperature, the place has high temperature throughout the years.

Table 4.17 The component transformation matrix for Equamax

Component	1	2
1	0.8520	0.5236
2	-0.5236	0.8520

Table 4.17 shows the component transformation matrix. This is used to see whether the rotation technique is suitable or not. Usually, a suitable rotation technique will result in a nearly symmetrical off-diagonal which is not true in the case.

Table 4.18: Component Score Coefficient Matrix (Eigen Vectors) for Equamax

	Component	
	1	2
Solar_Rad	0.4161	0.0320
RH	-0.5195	0.2717
Temp	0.2265	0.3764
WS	-0.2015	0.7118

Table 4.18 shows the eigenvectors of the two eigenvalues. The values represent the coefficient of the basis vectors. The first component shows that solar radiation and temperature have positive values, while RH and wind speed have negative values. From the second component, it can be seen that all have positive coefficients. However, wind speed, temperature and RH have larger positive coefficients with very small positive value in solar radiation.

iii) Oblique Rotation

In oblique rotation, the steps for extraction are as follows:

a. The variables are assessed for the unique relationship between each factor and the variables (removing relationships that are shared by multiple factors)

b. The matrix of unique relationships is called the pattern matrix.

c. The pattern matrix is treated like the loading matrix in orthogonal rotation.

When an oblique rotation is performed, two different matrices that can be used for interpretation are obtained: the pattern structure and factor correlation matrices.

If an oblique rotation, in which both spatial and temporal orthogonality are relaxed, is performed, the structures can be made even more close to zero and one, but the basic structure remains the same. The pattern matrix holds the beta weights to reproduce variable scores from factor scores.

There is a considerable disagreement about which of the following is the better basis for factor interpretation:

(i) Those who like using the structure matrix point out the long history of naming or interpreting factors in terms of the “variables with which they correlate.”

(ii) Those who like using the pattern matrix point out that there is often "simpler structure" in the pattern matrix

(iii) Those who like using the structure matrix point out that the apparent "simpler structure" (i.e., fewer multivocal items) in the pattern matrix is an illusion, made possible because of the correction for collinearity by the beta weights.

(iv) Typically, the interpretation based on the two matrices will be similar.

a) **Oblimin**: Tends to produce varimax-looking factors, but which are oblique.

Delta is a parameter that "controls" the extent of obliqueness amongst the factors.

i) Negative values "decrease" factor correlations

ii) "0" is the default

iii) Positive values (don't go over .8) "permit" additional factor correlation

Table 4.19: Total Variance Explained for Oblimin

Component	Initial Eigenvalues			Rotation Sums of Squared Loadings
	Total	% Var	Cum %	Total
1	2.3774	59.4342	59.4342	2.1746
2	1.1941	29.8529	89.2871	1.5830
3	0.2921	7.3021	96.5892	
4	0.1364	3.4108	100	

From Table 4.19, the last part titled: "Rotated Sums of Squared Loadings" gives the information for extracted components after rotation. Note that after rotation, only the relative value of eigenvalues has changed, the cumulative percentage remains the same.

The first principal component (scaled eigenvector) which explains the largest part of the total variance has a variance (eigenvalue) of 2.17. The second principal component has a variance 1.58. The first two principal components account for 89.29% of the total variance.

Note that after rotation only the relative value of eigenvalues has changed, the cumulative percentage remains the same.

Table 4.20: Pattern Matrix for Oblimin

	Component	
	1	2
RH	-0.9455	0.2561
Solar_Rad	0.8683	0.1609
Temp	0.6322	0.6139
WS	-0.0974	0.9883

From Table 4.20, this pattern matrix is related to correlating matrix in orthogonal rotation. In component 1, there is a large inverse relation of RH and good proportion loadings of solar radiation and temperature. Component 2 has high loadings of temperature and RH.

The rotated component for the Oblimin rotation represents both how the variables are weighted for each factor and the correlation between the variables and the components. Based on these components magnitude of the coefficients, it can be said that the first component represents dry season, and the second component represents rainy season. Because of the complex loadings of temperature, the place has high temperature throughout the years.

Table 4.21 Structure Matrix for Oblimin

	Component	
	1	2
RH	-0.8988	0.0833
Solar_Rad	0.8977	0.3196
Temp	0.7444	0.7294
WS	0.0832	0.9705

Table 4.21 indicates that the structure matrix holds the correlations between each variable and each factor (same as with orthogonal rotations) gives the components loadings after the rotation was carried out. For each of the variables, we get a loading in each of the columns representing factors. The variables are listed in the decreasing order of factor loadings as we requested the same in the Options window.

In component 1, there is a large inverse relation of RH and good proportion loadings of solar radiation and temperature. This is typical of the dry season. Component 2 has high loadings of temperature and wind speed and, by the nature of this place this is not typical of rainy season. Because of the complex loadings of temperature, the place has high temperature throughout the years.

Table 4.22: Component Correlation Matrix for Oblimin

Component	1	2
1	1	0.1827
2	0.1827	1

Table 4.22 shows the component transformation matrix. This is used to see whether the rotation technique is suitable or not. Usually, a suitable rotation technique will result in a nearly symmetrical off-diagonal which is realized in this case.

Table 4.23: Component Score Coefficient Matrix (Eigen Vectors) for Oblimin

	Component	
	1	**2**
Solar_Rad	**0.4169**	**0.0567**
RH	**-0.4818**	**0.2403**
Temp	**0.2715**	**0.3892**
WS	**-0.1117**	**0.6985**

Table 4.23 shows the eigenvectors of the two eigenvalues displayed in the Table. The values represent the coefficient of the basis vectors or weights of the variable.

b) ***Promax***

Table 4.24: Total Variance Explained for Promax

Component	**Initial Eigenvalues**			**Rotation Sums of Squared Loadings**
	Total	% of Variance	Cumulative %	Total
1	2.3774	59.4342	59.4342	2.2111
2	1.1941	29.8529	89.2871	1.7005
3	0.2921	7.3021	96.5892	
4	0.1364	3.4108	100	

From Table 4.24 the last part titled: “Rotated Sums of Squared Loadings” gives the information for extracted components after rotation.

The first principal component (scaled eigenvector), which explains the largest part of the total variance, has a variance (eigenvalue) of 2.21. The second principal component has a variance 1.70. The first two principal components account for 89.29% of the total variance.

Note that after rotation only the relative value of eigenvalues has changed as the cumulative percentage remains the same.

Table 4.25: *Pattern Matrix for Promax*

	Component	
	1	2
RH	-0.9861	0.3061
Solar_Rad	0.8627	0.1267
WS	-0.2057	1.0220
Temp	0.5743	0.6037

From Table 4.25, the pattern matrix holds the beta weights to reproduce variable scores from factor scores. After performing the oblique rotation, we obtain the correlation structure, in which we find that the first component has strong correlations with the three physical variables, Solar radiation, temperature and RH, and a weak correlation with wind speed, and the second component has strong correlations with wind speed and temperature.

The rotated component for the Promax rotation shows how the variables are weighted for each factor and the correlation between the variables, and the components. Based on these components loadings, it can be said that the first component represents dry season, and the second component represents rainy season. Because of the complex loadings of temperature, the place has high temperature throughout the years.

Table 4.26: *Structure Matrix for Promax*

	Component	
	1	2
Solar_Rad	0.9036	0.4053
RH	-0.8872	-0.0124
WS	0.1244	0.9556
Temp	0.7693	0.7892

Table 4.26 shows that the structure matrix holds the correlations between each variable and each factor (same as with orthogonal rotations) gives the components loadings after the rotation was carried out. For each of the variables, we get a loading in each of the columns representing factors.

After performing the oblique rotation we obtain the correlation structure, in which we find that the first component has strong correlations with the three physical variables (Solar radiation, temperature and RH), and a weak correlation with wind speed, and the second component has strong correlations with wind speed and temperature.

The rotated component for the Promax rotation shows how the variables are weighted for each factor and the correlation between the variables, and the components. Based on these components loadings, it can be said that the first component represents dry season, and the second component represents rainy season. Because of the complex loadings of temperature, the place has high temperature throughout the years.

Table 4.27: *Component Correlation Matrix for Promax*

Component	1	2
1	1	0.3230
2	0.3230	1

Table 4.27 shows the component transformation matrix. This is used to see whether the rotation technique is suitable or not. Usually, a suitable rotation technique will result in a nearly symmetrical off-diagonal which is realized in this case.

Table 4.28: *Component Score Coefficient Matrix (Eigen Vectors) for Promax*

	Component	
	1	2
Solar_Rad	0.4157	0.0987
RH	-0.4671	0.1843
Temp	0.2859	0.4080
WS	-0.0805	0.6699

Table 4.28 shows the eigenvectors of the two eigenvalues. The values represent the coefficient of the basis vectors or weights of the variables.

CONCLUSION

Based on the observations of all the component matrices, structure matrices and pattern matrices, it can be concluded that the area has two seasons: dry and rainy seasons. Based on the eigenvalues, it shows that dry season is longer than the rainy season, and this is typical to the area that has seven (7) months (58.3%) for the dry season and five (5) (41.7%) months for the rainy season. The complex loadings of temperature in the two components show that the area is very warm.

Based on the component transformation matrices for the orthogonal rotations and components correlation matrices for oblique rotation, it is clear that oblimin and promax are the most suitable rotations, due to the symmetry of the off-diagonal elements.

AKNOWLEDGEMENTS

First and foremost, I thank Allah for keeping me alive and making it possible for me to present this inaugural lecture. Next, I wish to appreciate the support of the management of Bayero University, Kano for sponsoring all my postgraduate studies and conferences throughout my stay in the University. I also appreciate the management of Bayero University, ably led by the Vice Chancellor Professor Muhammad Yahuza Bello, for the opportunity given to me to present this inaugural lecture.

Equally, I will forever remain grateful to my supervisor and mentor Prof. John T. Ndefru, who supervised my B Sc., M.Sc. and PhD (Physics), and Prof. Ado Dan'isa, who supervised my M.Eng. (Electrical) dissertation.

Special appreciation goes to my entire family members who have been very prayerful and supportive throughout my academic sojourn. I would also want to acknowledge the support of all the staff of Physics Department, Bayero University, Kano.

Last but not the least, I would also want to acknowledge the contributions of the two of my PhD students, in the persons of Rakiya Aliyu, of Kano University of Science and Technology Wudil, Kano and Sharafa Salihu Bolaji of Usman Danfodio University, Sokoto. These students are very hard working and are good research materials.

REFERENCES

Akgun, B. (2003). *Identification of periodic autoregressive moving average models.* Middle East Technical University.

Ångström A. K. (1929). On the atmospheric transmission of sun radiation and on dust in the air. *Geogr. Ann., 11*, 156-166.

ASA News (American Statistical Association News), (2016). American Statistical Association releases statement on statistical significance and p-values. provides principles to improve the conduct and interpretation of quantitative science. http://amstat.tandfonline.com/doi/abs/10.1080/00031305.2016.1154108#.Vt2XIOaE2MN .

Barnston A. G. & Livezey B. E. (1987). Classification, seasonality, and persistence of low-frequency atmospheric circulation patterns. *Monthly Weather Review 115*: 1083–1126.

Bell, W. R. (1984). An introduction to forecasting with time series models. Insurance: Mathematics and Economics 3, pp. 241-255.

Biggar, S. F., Gellman D. I., & Slater P. N., (1990). Improved evaluation of optical depth components from Langley plot data. *Remote Sens. Environ., 32*, 91-101.

Box G.E.P., Jenkins G., (1970). *Time series analysis, forecasting and control*, Holden-Day, San Francisco, CA.

Box, G. E. P., Jenkins, G. M. & Reinsel, G. C. (1994). *Time series analysis: Forecasting and control. Third Edition*. Prentice Hall.

Bruegge, C. J., R. N. Halthore, B. Markham, M. Spanner, & R. W. Wrigley, (1992). *Aerosol optical depth retrievals over the Konza prairie. J. Geophys. Res., 97,* 18743-18758.

Cattell, R. B. (1978). *The scientific use of factor analysis in behavioural and life sciences*. New York, NY: Plenum Press

Cheng X., Nitsche G., Wallace J. M. (1995). Robustness of low-frequency circulation patterns derived from EOF and rotated EOF analyses. *Journal of Climate 8:* 1709–1713.

Deepak, A., & Gerber H. E., (Eds.). (1983). Report of the experts meeting on aerosols and their climatic effects. *WCP-55, 107* pp. [Available from World Meteorological Organization, Case Postale No. 5, CH-1211 Geneva, Switzerland.]

Eck, T. F., Holben, B. N., Dubovic, O., Smirnov, A., Slutsker, I., Lobert, J. M., & Ramanathan, V. (2001a). Column-integrated aerosol optical properties over the Maldives during the northeast mon-soon for 1998–2000. *J. Geophys. Res., 106*, 28 555–28 566.

Eck, T. F., Holben, B. N., Reid, J. S., Dubovic, O., Smirnov, A., O'Neil, N. T., Slutsker, I., & Kinne, S. (1999). Wavelength dependence of the optical depth of biomass burning, urban, and desert dust aerosols. *J. Geophys. Res., 104*(D24), 31 333–31 349.

Eck, T. F., Holben, B. N., Ward, D. E., Dubovic, O., Reid, J. S., Smirnov, A., Mukelabai, M. M., Hsu, N. C., O' Neil, N. T., & Slutsker, I. (2001b). Characterization of the optical properties of biomass burning aerosols in Zambia during the 1997 ZIBBEE field campaign. *J. Geophys. Res., 106*(D4), 3425–3448,

Everitt, B. S. & Dunn, G. (2001). *Applied multivariate data analysis (2nd ed).* London: Arnold.

Fernandez M. G. (1995). Principal component analysis of precipitation and rainfall regionalization in Spain. *Theoretical and Applied Climatology. 50:* 169–183.

Fukuoka A. (1951). A study of 10-day forecast (A synthetic report). *The Geophysical Magazine, Vol. XXII*, Tokyo: 177–218.

Galadanci, G. S. M. & Tijjani, B. I. (2014). Effect of relative humidity on arctic aerosols. *Advances in Physics Theories and Applications Vol. 36,* 2014; ISSN 2224-719X (Paper); ISSN 2225-0638 (Online). *www.iiste.org*

Hasmida, H. (2009). *Water quality trend at the upper part of johor river in relation to rainfall and run-off pattern*. Universiti Teknologi, Malaysia.

Hendranata, A. (2003). *ARIMA (Autoregressive moving average)*. Manajemen Keuangan Sektor Publik FEUI.

Hess M., Koepke P., & Schult I. (1998). Optical properties of aerosols and clouds: The software package OPAC, *Bulletin of the American Meteorological Society, Vol. 79,* No. 5, 831-844.

Ho, S. L. & Xie, M. (1998). The use of ARIMA models for reliability forecasting and analysis. *Computers ind. Engng, Vol. 35*, Nos 1-2, pp. 213-216.

Horel J. D. (1981). A rotated principal component analysis of the interannual variability of the northern hemisphere 500 mb height field. *Monthly Weather Review 109:* 2080–2092.

Iqbal M., (1983). *An introduction to solar radiation*. Academic Press.

Jolliffe I. T. (1987). Rotation of principal components: Some comments. *Journal of Climatology 7:* 507–510.

Junge, C. E., (1963). *Air chemistry and radiochemistry*. Academic Press.

Kasten, F. (1968). Der einflu5 der aerosol-grobenverteilung und ihrer hderung mit der relativen feuchte auf die sichtweite. *Beitrcige zur Phyaik der Atmosphfire 41,* 33-61.

Kaufman, Y. J. (1993). Aerosol optical thickness and atmospheric path radiance, J. *Geophys. Res., 98,* 2677-2992.

Kawamura R. (1994). A rotated EOF analysis of global sea surface temperature variability with inter-annual and inter-decadal scales. *Journal of Physical Oceanography 24:* 707–715.

King, M. D. & Byrne, D. M. (1976). A method for inferring total ozone content from spectral variation of total optical depth obtained with a solar radiometer. *J. Atmos. Sci., 33,* 2242–2251.

Koschmieder, M. (1926). *Theorie der horizontalen sichtweite. Beitrcige zur Physik der frekn Atmosphiire 12,* 33-66 and 171-181.

Lee, C. and Ko, C. (2011). Short-term load forecasting using lifting scheme and ARIMA models. *Expert Systems with Applications, Vol. 38,* pp. 5902-5911.

Lorenz E. N. (1956). *Empirical* orthogonal functions and statistical weather prediction. *Technical Report, Statistical Forecast Project.*

Lorenz E. N. (1970). Climate change as a mathematical problem. *Journal of Applied Meteorology 9:* 325–329.

Mestas-Nunez A. M. & Enfield D. B. (1999). Rotated global modes of non-ENSO sea surface temperature variability. *Journal of Climate 12:* 2734–2746.

Middleton, W. E. K. (1962). *Variation through the atmosphere*. Toronto, University of Toronto Press, pp. 104-106.

Montgomery, D. C. & Johnson, L. A. (1976). *Forecasting and time series analysis.* New York: McGraw-Hill.San Francisco, CA: Holden-Day.

Montgomery, D. C., Jennings, C. L. & Kulahci, M. (2008). *Introduction to time series analysis and forecasting*. John Wiley & Sons, Inc.

Moorthy K. K., Saha A., Prasad B. S. N., Niranjan K., Jhurry D.& Pillai P. S. (2001). Aerosol optical over peninsular India and adjoining oceans during the INDOEX campaigns: Spatial, temporal and spectral characteristics; *J. Geophys. Res.10628*,539–28,554.

O'Neill, N. T., Dubovic, O., & Eck, T. F.(2001). Modified Angstrom exponent for the characterization of submicrometer aerosols, *Appl. Opt., 40*(15), 2368–2375.

O'Neill, N. T., Eck, T. F., Smirnov, A., Holben, B. N., & Thulasiraman, S. (2003). Spectral discrimination of coarse and fine mode optical depth. *J. Geophys. Res., 198*(D17), 4559, doi:101029/2002JD002975.

Richman M. B. (1986). Rotation of principal components. *Journal of Climatology 6:* 293–335.

Richman M. B. (1987). Rotation of principal components: A reply. *Journal of Climatology 7:* 511–520.

Satheesh S. K. & Moorthy K. K. (1997). *Aerosol characteristics over coastal regions of the Arabian Sea*; Tellus 49B, 417–428.

Shaw G. E., Regan J. A. & Herman B. M. (1973). Investigations of atmospheric extinctions using direct solar radiation measurements made with a multiple wavelength radiometer. *J. Appl. Meteor.*12374–380.

Singh S., Nath S., Kohli R. & Singh R. (2005). Aerosols over Delhi during pre-monsoon months: Characteristics and effects on surface radiation forcing. *Geophys. Res. Lett. 32*L13808 doi:10.1029/2005GL023062.

Smith, I. (1994). Indian Ocean sea-surface temperature patterns and Australian winter rainfall. *Int. J. Climatol., 14,* 287-305.

Thurstone L. L., (1947). *Multiple factor analysis*. The University of Chicago Press, Chicago, IL.

Tijjani B. I., Sha'aibu F. & Aliyu A. (2013). The effect of hygroscopic growth on marine aerosols. *Fire Journal of Natural and Applied sciences.1*(2),70-88.

Tijjani B. I., Sha'aibu F. & Aliyu A. (March 2014). *The effect of relative humidity on maritime polluted aerosol*s. *International Journal of Pure and Applied Physics Vol.2,* No.1, pp.9-36.