ISSN: 2454-1907
Impact Factor: 1.745 (I2OR)

IJETMR
International Journal of Engineering Technologies and Management Research
A knowledge Repository

APPLICATION OF CANONICAL CORRELATION ANALYSIS ON SCIENCE PRODUCTION

Abdulmuahymin A. Sanusi [1], Muhammad B. Muhammad [2]
[1, 2] Department of Mathematics and Computer Science, Federal University Kashere, Gombe state, NIGERIA
DOI: 10.5281/zenodo.61366

Abstract:
The development of any given nation is behind its advancement in science and technology; this can be achieved through the upbringing of the new generation on the knowledge related to science and technology by putting in all efforts, factors and mechanisms that will easily aid the better understanding and interest in science and technology. This research work tends to investigate the production of science in some selected schools in Gombe state, Nigeria; that offer science as their core subjects. Three factors were used; School output [i.e. the grades obtained in science subjects(Mathematics(MTH), Physics(PHY), Chemistry(CHM) and Biology(BIO)], the School input [i.e. Averagely Equipped Library and Laboratory for Science (AELAL), Science teachers' years of teaching experience (STYTE), Instructional Hours on Science subjects per week (INSHR) and students' teacher ratio (STR)] and Environmental input [i.e. The number of text books on science possessed by students (NTBS), hour spent studying science outside school hours (HRSS), home leaning aids on science such as computer, science dictionary est. (HLAS) and home extra moral teacher on science(HETS)]. Two sets were formed, Set-A (school output) and Set-B (school input and environmental input).The data used is obtained through the questionnaire distributed to the random selected school. The research work adopts the use of Descriptive statistics to verify the normality of the data and Canonical Correlation Analysis to investigate the relationship between the sets of the data. Three Canonical roots were obtained and only two are statistically significant, the first showing a strong positive correlation coefficient between the sets of data, indicating the impact of the School and Environmental inputs on the school output. However, improvement on the School and Environmental inputs will equally improve the production of Science in the selected schools as a case study and some other schools in the states at large.

Keywords:
Canonical Correlation, Production of Science, Science and Technology.

Cite This Article: Abdulmuahymin A. Sanusi, and Muhammad B. Muhammad, "APPLICATION OF CANONICAL CORRELATION ANALYSIS ON SCIENCE PRODUCTION" *International Journal of Engineering Technologies and Management Research*, Vol. 3, No. 8(2016)15-24.

ISSN: 2454-1907
Impact Factor: 1.745 (I2OR)

1. INTRODUCTION

Science production has been one of the outmost mechanism needed by every country to improve in the development of the aspect that affect their technology, through the production of various materials, equipment's and devices that are basically needed to achieve or asses huge amount of result within a twinkle of an eye. The reasons for science production in our present age are almost as complex as are the reasons we are unable to under determine in vast numbers.

In the world today, statistics show that the developed countries have gone far and deeply vast in the aspect of science and technology over many years. For the developing country in the aspect of science, the basic knowledge on science and those factors that will easily facilitate the production of science should be look into.

In Nigeria, there is no doubt that the global science developments crises have necessitated sudden changes in the mind of our local scientist in the recent times in order to prevent science recession. This has caused abrupt movement in the production of science and the growth of science. Hence, there is need to examine the impact of School inputs and Environmental inputs on Science Subjects in Nigeria. This is the thrust for this research study.

Science production in secondary schools/ high schools is the most important factors in the promotion of science capacity building of any country. It enables countries to build an indigenous science based on solid foundation. Consequently, an investigation on how school and environmental inputs into science production process affect science subjects. Furthermore, Hanushek (1979) noted that science professors found that students' performance in mathematics is correlated with their performance in science.

As outlined by O'Sullivan (2000), school achievement depends on five inputs: the school curriculum, educational equipment, the classroom teacher, the home environment, and the achievement level of the child's classmate. In general, these five inputs to the production function can be divided into three groups: school resources, environmental inputs and peer group effects. In this study, only the effects of school resources, environmental inputs and students' grades in science subjects are investigated. School inputs include; (Averagely Equipped Library and Laboratory for Science, Science teachers' years of teaching experience, Instructional Hours on Science subjects per week and students' teacher ratio). On the other hand, environmental inputs include (Number of text books on Science possessed by students, Hours spent for studying science outside the school hours, Home Learning aids on Science, Home extra moral teacher on science). And the school output is the students' grades in science subjects (Mathematics, Physic, Chemistry and Biology).

2. METHODOLOGY

2.1.DESCRIPTIVE STATISTICS

Basic descriptive statistics are calculated to 64 bit decimal precision avoiding any of the pocket calculator formulae that led to unnecessary lack of precision (McCullough and Wilson, 1999).

ISSN: 2454-1907
Impact Factor: 1.745 (I2OR)

Mean = $\overline{x} = \frac{\sum_{i=1}^{n}(x_i)}{n}$ (1.1)

Standard deviation = $S = \sqrt{\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1}}$ (1.2)

2.2.CANONICAL CORRELATION APPROACH MODEL

An initial step in canonical correlation analysis is an inspection of the correlation matrix of the given data.
Let S denote the data such that: S = {Set-A, Set-B}
Where:

Set – **A** = {**MTH**, **PHY, CHE, BIO**}
Set – **B** = {**AELAL, STYTE, INSHR, STR, NTBS, HRSS, HLAS, HETS**}

Proper analysis begins with a simple examination of the correlation significance Dunn et. al.(1977).

The research work proposed Canonical Correlation Analysis Approach for the analysis. Canonical correlation's goal is to quantify the strength of the relationship, in this case between the two sets of variables. Thus, canonical correlation identifies the optimum structure or the dimensionality of each variable set that maximizes the relationship between dependent and independent variable sets.

Canonical correlation analysis deals with the association between composites sets of multiple dependent and independent variables. In doing so, it develops a number independent canonical function that maximize the correlation between the linear composites, also known as canonical variates, which are sets of dependent and independent variables. . Among unique feature of canonical correlation is that the variates are derived to maximize their correlation. Moreover, canonical correlation does not stop with the derivation of a single relationship between the sets of variables, instead a number of canonical functions.

Canonical correlation analysis reduces each of these patterns to derived variables, the canonical U and V variables. The largest canonical correlation corresponds to the strongest relation between independent and dependent variables. Sub-sequent canonical correlations correspond to relation of decreasing strength. For example, different patterns of flight mode selection under different phases of flight Canonical correlation analysis allows these patterns to be character objectively and allows their relative strengths to be measured. Anderson (1958).
Anderson (1958) gave a detailed Mathematical concept of canonical correlation analysis. Let X be a q-dimensional random vector and Y be a p-dimensional random vector. Suppose that X and Y have means μ and ν respectively and that

$$\mathrm{E}\left[(x-\mu)(x-\mu)^{'}\right]=\Sigma_{11} \quad (2.1)$$

$$\mathrm{E}\left[(y-\nu)(y-\nu)^{'}\right]=\Sigma_{22} \quad (2.2)$$

ISSN: 2454-1907
Impact Factor: 1.745 (I2OR)

$$\mathrm{E}\left[\left(x-\mu\right)\left(y-\nu\right)^{'}\right]=\sum_{12}=\sum^{'}{}_{21} \tag{2.3}$$

Let us now consider the linear combinations

$$\mathrm{g}=a^{'}x \tag{2.4}$$

and

$$\mathrm{f}=b^{'}y \tag{2.5}$$

The correlation between g and f is defined as show below

$$\rho(a,b)=\frac{a^{'}\sum_{12}b}{\left[\left(a^{'}\sum_{11}a\right)\left(b^{'}\sum_{22}b\right)\right]^{1/2}} \tag{2.6}$$

Tests for Significance using Wilk's Lambda Test.

The Wilk's Lambda test of hypothesis is given as:

$H_o:\sum_{xy}=0$, i.e. there is no relationship between the canonical variates.

$H_1:\sum_{xy}\neq 0$, i.e. there is relationship between the canonical variates.

Test statistic:

$$\lambda_1=\frac{|R|}{|R_{yy}||R_{xx}|} \tag{2.7}$$

Where:

R is the correlation between $x^{'}s$ *and* $y^{'}s$

R_{xx} is the correlation between $x^{'}s$

R_{yy} is the correlation between $y^{'}s$

Significance Level:

$\alpha=0.05$

Decision rule:

Reject H_o if $p\prec 0.05$ and otherwise accept. Rencher (2002)

3. DATA USED FOR THE ANALYSIS

The data used for the Analysis is generated from Secondary schools that offer science subjects in Gombe State, random selection of twenty seven (27) schools were made of which five (5) students were also randomly selected from each school; through the questionnaires dispatched among these schools, required information about the students and the schools were generated. The questionnaire is structured to contain some expressions such as the school/educational output (i.e. the grades obtained in science subjects), the school input (i.e. averagely equipped Library and Laboratory on science, Science teachers' years of teaching experience on science, Instructional Hours spent on Science subjects per week, and Students' teacher ratio) and the

ISSN: 2454-1907
Impact Factor: 1.745 (I2OR)

environmental input (i.e. the number of text books on science possessed by the students, hour spent studying science outside school hours, home leaning aids on science such as computer, science dictionary est. and home extra moral teacher on science)

4. ANALYSIS AND DISCUSSION

Table 1; is the mean values and standard deviation of each variable considered in the analysis. It is not surprising that the mean scores for mathematics, Physics and Biology are around 60 since most of the students' grades in the subjects are B, while Biology is around 70, indicating the students' grades in the subject is A. The variables with the highest mean in this study is the Instructional Hours on science, Averagely equipped library and Laboratory on science and the Number of text books on science possessed by the students; suggesting most of the schools under this study had very high instructional hours on science, maintain averagely equipped library and laboratory on science and most of the students possess a meaningful text books on science.

Table 1: Descriptive statistics

Variable	Frequency	Mean	Standard deviation
School Output			
MTH	135	64.9741	14.0120
PHY	135	66.9926	14.9544
CHM	135	66.9407	12.7857
BIO	135	71.4778	13.7727
School Input			
AELAL	135	4.4519	1.9877
STYTE	135	3.5644	1.3412
INSHR	135	9.9430	1.6238
STR	135	0.6296	0.4847
Environmental Input			
NTBS	135	4.3111	1.4838
HRSS	135	2.6000	1.3505
HLAS	135	0.8222	0.3837
HETS	135	0.6074	0.4901

Table 2: Canonical correlation coefficient of Set – A and Set – B

Canonical Functions	Canonical Correlation	Eigen values	% of Variance Explained
1	0.5877	0.3453	59.7
2	0.3843	0.1477	25.5
3	0.2922	0.0853	14.8

Table 2 shows the Canonical correlation of the three canonical variates and their corresponding Eigen values. The Eigen values of the canonical variates can be tested by employing Wilk's Lambda criterion to test for the significant by using Wilk's Lambda test, Rencher (1998).

Hypothesis:

$H_o : \sum_{xy} = 0 \quad Against \quad H_1 : \sum_{xy} \neq 0 \quad at \ \alpha = 0.05$

Reject H_o *if* $p \prec \alpha = 0.05$, we have the following table:

Table 3: Shows the Wilk's Lambda test

S/NO	**N**	P	Q	Df	p-value	$\alpha - value$
1	135	8	4	32	0.0000	0.05
2	135	7	3	21	0.0098	0.05
3	135	6	2	12	0.0979	0.05

From table 3 above, the canonical correlations tested is significant at the first and second canonical correlation coefficient with p_1 – value = 0.0000 and p_2 – value = 0.0095 $\prec \alpha = 0.05$, since the p-values of the first two canonical variate are less than the alpha value, it implies that the null hypothesis is rejected. This indicates that two of the three canonical correlation coefficients are significantly different from zero. **'P'** is the number of variables considered in a certain canonical variate, while **'Q'** is the number of variables considered in the opposite canonical variate and **'df'** is the degree of freedom used at each level of canonical function.

We therefore consider the first canonical variate pair U_1 and V_1 with canonical correlation Coefficient r_1= 0.5877 as it significant and possess the highest degree of canonical correlation coefficient, so that the proportion of variance common to the first canonical variate pair is $r_1^2 = 0.3453$ showing about 34.53% of the proportion of variance captured by the first canonical variate.

Similarly r_2= 0.3843 is the canonical correlation coefficient between the second canonical variate pair and so $r_2^2 = 0.1477$ which indicates about 14.77% of the proportion of variance captured, r_3= 0.2922 shows the canonical correlation coefficient between the third canonical variate pair and so $r_3^2 = 0.0853$ indicating 8.53% of the proportion of variance captured.

Table 4: Canonical loading for Set –A and Set – B

ISSN: 2454-1907
Impact Factor: 1.745 (I2OR)

Sets	Variables	r_1	r_2	r_3
Set – A	***School Output***			
	MTH	0.4715	0.6337	-0.5651
	PHY	0.3209	0.1523	1.1496
	CHM	0.5364	0.5924	-0.5359
	BIO	-0.0736	-0.6003	0.0004
Set – B	***School Input***			
	AELAL	0.4645	-0.7003	-0.0919
	STYTE	-0.6165	-0.5731	-0.0466
	INSHR	0.6066	-0.0885	-0.7483
	STR	0.2204	-0.3887	-0.0646
	Environmental Input			
	NTBS	0.5036	-0.0366	0.5828
	HRSS	0.2578	0.3565	-0.0136
	HLAS	-0.4606	-0.1713	0.2146
	HETS	-0.1133	-0.3679	-0.0342

Table 4: A canonical loadings that provide information about the relative contribution of variables to each independent canonical relationship, the first pair of canonical variates can be written as follows:

U_1 = 0.4715MTH + 0.3209PHY + 0.5364CHM – 0.0736BIO

V_1 = 0.4645AELAL - 0.6165STYTE + 0.6066INSHR + 0.2204STR + 0.5036NTBS + 0.2578HRSS – 0.4606HLAS - 0.1133HETS

$\emptyset$ =0.5877
The correlation $\varnothing_1$ between U_1 *and* V_1 is called the first canonical correlation coefficient.

Looking at the contribution of the individual variable used in the analysis irrespective of the negative signs, in Set-A; CHM is loading the heaviest value 0.5364, followed by MTH (0.4715), PHY (0.3209) and Biology (0.0736), while in Set-B; STYTE loading heaviest with the value (0.6165), followed by INSHR (0.6066), and NTBS (0.5036), while other variables loadings such as AELAL (0.4645), HLAS (0.4606), HRSS (0.2578), STR (0.2204) and HETS (0.1133) are values less than 0.5 indicating their lower contribution and impact to the first canonical coefficient.

Thus, the values attached to each variable in Set-A and Set-B are their partial correlation to their corresponding canonical variables and indicating the individual contribution to the canonical pair.

Table 5: Canonical cross loading for Set-A and Set-B

Sets	Variables	r_1	r_2	r_3
Set-A	***School Output***			
	MTH	0.4465	-0.2101	-0.0412
	PHY	0.4323	-0.0265	0.1968
	CHM	0.4567	-0.1645	-0.0795
	BIO	0.0892	-0.2628	0.0149
Set-B	***School Input***			
	AELAL	0.1085	-0.2374	0.0794
	STYTE	0.2361	-0.2007	-0.0505
	INSHR	0.1247	-0.0583	-0.2284
	STR	0.2853	-0.0616	-0.0293
	Environmental Input	0.2629	-0.0922	0.1910
	NTBS	0.1750	0.0646	0.0440
	HRSS	0.1782	-0.1241	0.0646
	HLAS	0.0563	-0.0714	0.0069
	HETS			

Table 5 shows the Canonical Cross loading of the three canonical functions. In the first canonical function, Set A, it can be seen that CHM, MTH and PHY slightly have almost average correlations with independent canonical variate 0.4567, 0.4465 and 0.4323 respectively while BIO with a very weak correlation 0.0892, from Set-B, i.e. STR with 0.2853 followed by NTBS with 0.2629, followed by STYTE with 0.2361, followed by HLAS with 0.1782 up to the last variable with the weakest correlation, that is HETS with 0.0563.

However, the canonical correlation which examines the linear relationship between Set – A and Set – B variables is by creating the combinations. The first canonical correlation explains the maximum relationship between the canonical variates and each successive canonical correlation is estimated so as to be orthogonal yet still explain the maximum relationship not accounted for by the previous canonical correlation. This reflects the high variance among these variables. By squaring the terms in the canonical loading, we find percentage of the variance for each of the variable explained by function 1.

5. CONCLUSION AND RECOMMENDATION

We observe that set-A and set-B are strongly correlated at the first canonical correlation variate. However, Canonical correlation analysis measured the strength of relationship of the canonical pair and the variables that strongly contributed. The first pair with a measure of correlation of 0.5877 with the proportion of variability of about 59.7%, the second pair with a measure of correlation 0.3843 with the proportion of variability of about 25.5% and the third canonical pair with a measure of correlation 0.2922 having a proportion of variability of about 14.8%.

From the output of the analysis carried out on the entire data, it is apparent that the correlation between Set-A (School output) and Set-B (School input and Environmental input) is a strong positive correlation at the first canonical variate due to strong contribution of Science teachers'

years of teaching experience, Instructional hours on science (School input) and Number of text books on science possess by students (Environmental input). While averagely equipped Library and Laboratory on science, Home learning aids on science, Hours spent studying science outside the school, Students teacher ratio and Home Extra moral teacher on science contribute weakly.

It is recommended that all other variables that contribute weakly to the production of science such as School input (averagely equipped Library and Laboratory on science, Students teacher ratio) and Environmental input(Home learning aids on science, Hours spent studying science outside the school and Home Extra moral teacher on science) should be improved and encouraged in schools and at home respectively to boost the production of science in our society, this will equally encourage all science students to think towards what he/she can provide and produce through the little knowledge acquired on science and eventually brings about development of the nation in terms of science and technology as it is obtainable in some developed nations such as China, Japan, Saudi Arabia, America among others.

6. REFERENCES

[1] *Abdulmuahymin et. al. (2016): Analysis of Government Policy on the Youth in the Causes of Violence Using Canonical Correlation Analysis. Sub-Sahara African Journal of Humanities and Social Science, 3(4): 191-201.*

[2] *Anderson, T. W. (1958): An Introduction to Multivariate statistical Analysis. First Edition, John Wiley and sons. New York.*

[3] *Anderson R. L. Tathan and Willian C. Bank (1998): Multivariate Data Analysis,* 5^{th} *Edition Prentic Hall, New York.*

[4] *Borga, M. (2001): CanonSical correlation, A tutorial, http://people. Imt liu. se/~magnis/ccal, unpublished*

[5] *Claude Ake (2002), A Political Economy of Africa, Nigeria: Longman Nigeria.*

[6] *Davatzikos, C. (2004): Predict Modelling of Anatomic structure, using canonical correlation Analysis. University of Pennsylvania, Philadelphia.*

[7] *Dunn J.W.; and Doekson,G.A. (1977): Canonical correlation analysis of selected Demographic and Health Personnel Variables. Southern Journal of Agricultural Economics. 2: 565-570.*

[8] *Everitt, B.S.; and Dunn G. (1991): Applied Multivariate Data Analysis. Edward Arnoid. London. Pp 219-220.*

[9] *Hair, J.F., Anderson, R.E, Tathaman, R.L, & Black,W.C (1998): Multivariate data Analysis, Fifth edition. NJ. Prentice Hall.*

[10] *Hallow, L.L (2005): The Essence of Multivariate Thinking, Theme and methods. Mahwah, N.J; Lawrence Erlbaum Associates, pages 205,229.*

[11] *Hotelling, H. (1936): Relations Between two sets of Variable. Biometrika. 28:312-377Liu. T.; Shen, D, and Marada K.V.; Kent J. T.; and Bibby J.M. (1979): Multivariate Analysis, fifth edition Academic press inc, London. Okwudiba Nnoli (1978), Ethnic Politics in Nigeria, Enugu: Fourth Dimension publishers.*

[12] *Rencher, A.C (2002): Methods of Multivariate Analysis. Second edition, John Wiley & Sons. Inc. New York.*

[13] Singh et. al. (2013). Analysis of Science students' SSCE results using Canonical Correlation Analysis. Continental Journal Education Research. 6(1): 22 – 26.

[14] Schul P. L.; William M.P.; and Taylor L. (1983): The Impact of Channel Leadership Behavior on Intra channel Conflict. Journal of Marketing. 47(3): 21-34.

[15] Shafto, M.G.; Degani A. and Kirlik, A. (1997): Canonical Correlation Analysis of Data on Human-Automation Interaction, proceeding of the 41^{st} Annual meeting of the Human Factors and Economics Society Albuquerque. NM, Human Factor Society.

[16] Simon, H. A (1969): The Science of Artificial, Cambridge M.A., M.I.T Press. London.

[17] Thompson, B. (1984): Canonical Correlation Analysis: Uses and interpretation series: Quantitative applications in the social sciences. 47: 1-71.

[18] Van Auken, E. (1993): A Financial comparison Between Korean and U.S, A Cross Balance sheet Canonical Correlation Analysis. Journal of small business Management. 31(3): 73-83.