Vol.6.Issue.3.2018 (July-Sept.)
©KY PUBLICATIONS

http://www.bomsr.com
Email:editorbomsr@gmail.com

RESEARCH ARTICLE

BULLETIN OF MATHEMATICS AND STATISTICS RESEARCH

A Peer Reviewed International Research Journal

INTERNATIONAL STANDARD SERIAL NUMBER 2348-0580

MULTIPLE DISCRIMINANTANALYSIS AS APPLIED TOLANGUAGE DISTINCTION

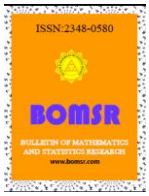
SHUKURANI MAGUNDO[1], GEORGE MUHUA[2], ERICK OKUTO[3]

[1]Department Of Mathematics, Catholic University of Eastern Africa, P.O Box 62157-00200, Nairobi, Kenya

[2]School Of Mathematics, University of Nairobi, P.O Box 30197, GPO, Nairobi, Kenya

[3]Department Of Applied Statistics, JaramogiOgingaOdinga University of Science & Technology, P.O Box 210-40601, Bondo, Kenya.

Correspondece author (ShukuraniMagundo: magundo2013@gmail.com)



ABSTRACT

This paper presents study on the application of multiple discriminant analysis (MDA) to distinguish between languages with a focus on five languages of the Coastal region of Kenya. Chapter one gives an introduction of the paper, chapter two explains the methodology used, chapter three presents the results, chapter four gives a brief discussion of the findings, and lastly chapter five presents the conclusions and recommendations.

Key words:Intelligibility, Lexicon, Gender, Multiple Discriminant Analysis

1.0 Introduction

Sociolinguists have tried to distinguish between languages the world over. Some of them have based their studies on the frequencies of occurrences of linguistic elements (Gries, 2015). Several others have used the comparison of means and percentages of linguistic similarities. These include: Faton(2018) in Togo/Benin, Webster (2017) in India/Nepal, Beine(2017) in central India, Lambrecht and Mann (2017) in Cambodia, Kassell, MacKenzie, and Potter (2017) in Papua New Guinea, Jordan and Manuel (2016) in Angola, Muniru, Magnusson, Hansley, and Ayenajeh(2016) in Nigeria, Lebold and Young Lee (2016) in Indonesia, Gonzales (2015)in Philippines, and lastly Kröger(2005)in Tanzania and Mozambique. In another study, Kluge (2008) employed the comparison of means and standard deviations. Although these methods have worked well, they have employed univariate techniques and the researchers have had to look at each variable singly, sometimes making it difficult to decide whether two or more languages in an area are different.

This study then applies the use of the multiple discriminant analysis (MDA), analyses several variables together, reduces errors and thus brings about precision in decision making.

2.0 Methodology

Discriminant analysis is a technique that is used to analyse data where the dependent variable is categorical and the independent variables are numerical in nature. If the dependent variable has two groups then it is just called Discriminant Analysis or Fisher's Linear Discriminant analysis (LDA). However, if the dependent variable has three or more groups then it is referred to as Multiple Discriminant Analysis (MDA) or Canonical Discriminant Analysis (CDA). Discriminant analysis aims at developing discriminant functions or linear combinations of independent variables that will discriminate between groups in the dependent variables. The weights of each of the independent variable are known as the discriminant coefficients(Statistics Solutions, 2018; Rencher, 2002). The objective of multiple discriminant analysis includes examining group separation (profiling) in a two dimensional plot (Rencher, 2002), ranking variables in terms of their relative importance to the separation of groups (Rencher, 2002) and determining group membership of samples from a group of predictors by finding linear combinations of the variables which maximize the differences between the populations being studied, with the objective of establishing a model to sort objects into their appropriate populations with minimal error" (Brown, 1998). Since the main aim of the research was to describe group differences (also known as profiling) descriptive multiple discriminant analysis was used as opposed to predictive discriminant analysis whose main purpose is to classify subjects into one or several known groups (or classifying).

2.1 MDA as applied in this study

Let X_i denote *p=3* independent random variables and *n=200* be the number of observations for each independent variable. Let *k=30* samples be taken from each of the five *(g=5)*language populations. We obtain a multiple discriminant function of the form,

$$\mathbf{Z} = a_1\mathbf{X} + a_2\mathbf{X} + a_3X = \mathbf{a'X} \quad \ldots \quad (2.1)$$

Where the vectors, $\mathbf{a}_i$, which form the matrix **a** are the eigenvectors of $\mathbf{E^{-1}H}$ corresponding to the eigenvalues λ_i, i = 1, 2, 3 and $\mathbf{X} = (X_1, X_2, X_3)$ is a vector of variables. This gives the minimum of *(g-1),p* multiple discriminant functions.

The matrix of coefficients, **a**, is obtained from

$$(\mathbf{E^{-1}H} - \lambda\mathbf{I})\mathbf{a} = 0 \quad \ldots \quad (2.2)$$

Where,

E is the within-group sums of squares matrix,

and

H is the between-group sums of squares matrix.

λ_i is the eigenvalue corresponding to variable X_i

2.2 Contribution to group separation

Using the MDA function that best separates the groups, the discriminant function coefficients, a_{ij}, are assessed in order to determine the relative importance of the contribution of each variable, X_i, to the separation of the groups. Since the measurements in this study are commensurate, the absolute values of corresponding coefficients are used in order to assess the importance of each variable to the separation of the groups.

2.3 Variables and sample size

Three variables are considered namely Lexicon (lexical similarity), Intelligibility and Gender. Lexicon is determined by first collecting a wordlist of 200 words from elderly people from the Giryama speech community. Then a wordlist of 200 words is collected from 30 individuals in each of the five speech communities. Each individual's wordlist isthen marked based on the wordlist collected from the Giryama elders.

Intelligibility is determined by administering an intelligibility test based on Blair (1997) and Casad(1974), amongst 30 individuals in each of the five speech communities. The intelligibility test is administered using a Giryamastory.Briefly, the variables are X_1 = Lexicon (lexical similarity), X_2 = Intelligibility, and X_3 = Gender. Simple random selection is used to select the 30 individuals in each of the five speech communities.

3.0 Results

3.1 Matrix scatterplot of raw data

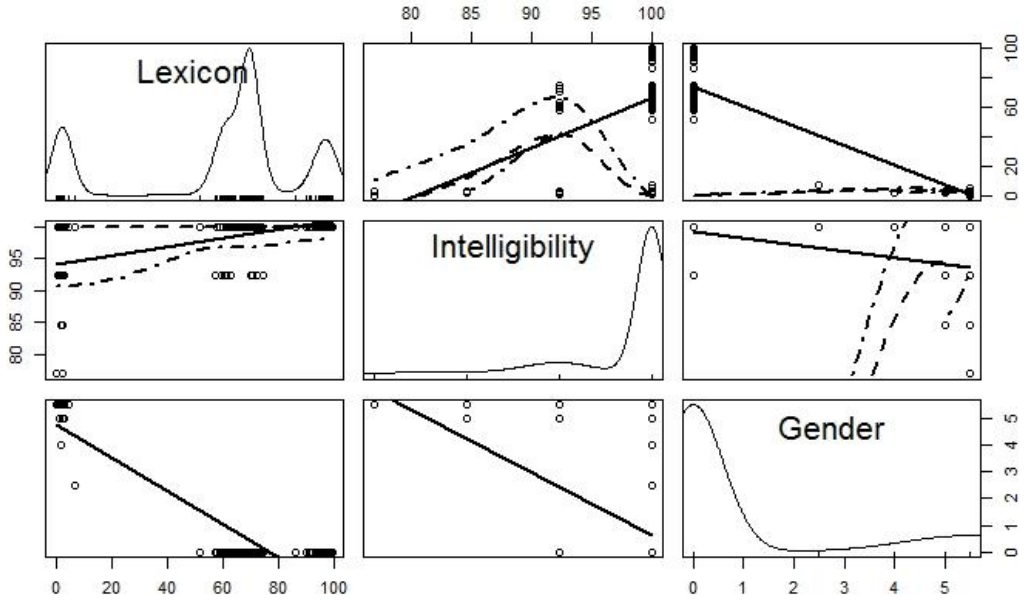


Figure 1: Shows a matrix of scatter ofthe five languages

Figure 1 above is a graphical representation of the raw data collected. The matrix is symmetrical hence, cells in the upper triangle are similar to their corresponding cells in the lower triangle.

3.2 Group means

Table 1:Group means

	LEXICON	INTELLIGIBLITY	GENDER
CHONYI	69.533333	99.23000	0.000000
DURUMA	60.133333	98.46000	0.000000
GIRYAMA	95.983333	100.00000	0.000000
JIBANA	69.933333	98.97333	0.000000
WAATA	2.416667	93.84000	5.283333

Table 1 above is a summary of the group means. It shows the mean of each variable in every language group.

3.3 Contribution to group separation

Note: LD1 = 1st Discriminant Analysis function, LD2 = 2nd Discriminant Analysis function, and LD3 = 3rd Discriminant Analysis function

Table 2: Coefficients of multiple discriminants

	LD1	LD2	LD3
LEXICON	-0.33530441	-0.21786462	-0.01027808
INTELLIGIBILITY	0.01397462	-0.02100527	0.25561652
GENDER	1.61409210	-3.34909990	0.11861214

Table 2 represents the coefficients of the multiple discriminant functions and it forms the matrix **a'**(**a**transpose).

From the formula Z = **a'**X,where X is the vector of variables, we obtain the following discriminant functions.

Z_1 = -0.33530441Lexicon + 0.01397462Intelligibility
+ 1.61409210Gender

Z_2 = -0.21786462Lexicon -0.02100527Intelligibility
-3.34909990Gender

Z_3 = -0.01027808Lexicon + 0.25561652Intelligibility
+ 0.11861214Gender

3.4 Relative importance of each discriminant function

Table 3:Proportions of trace

LD1	LD2	LD3
0.9611	0.0389	0.0000

Table 3 above gives the proportions of trace. Each proportion of then determines the relative importance of each discriminant function.

3.5 A plot of the discriminant functions

In order to visualise the way the discriminant functions separate the groups, graphs are plotted representing two pairs of the discriminant functions at a time.

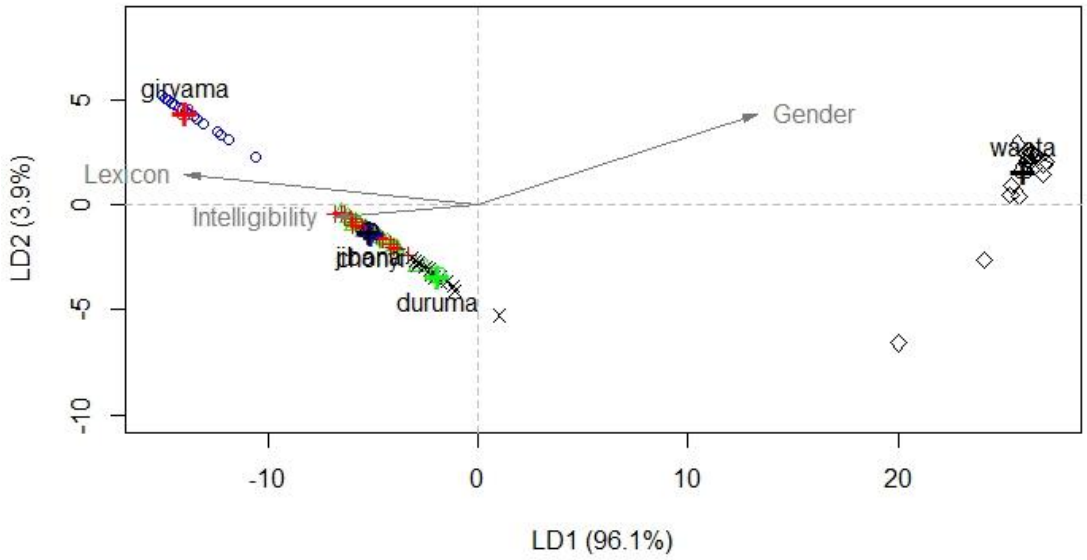


Figure 2: LD1 vs LD2

Upon plotting the first two discriminant functions LD1 and LD2, as represented in Figure 2 above we see how the points are organized and grouped after the within-groups variances are minimized and the between-groups variances are maximized by the discriminant functions.

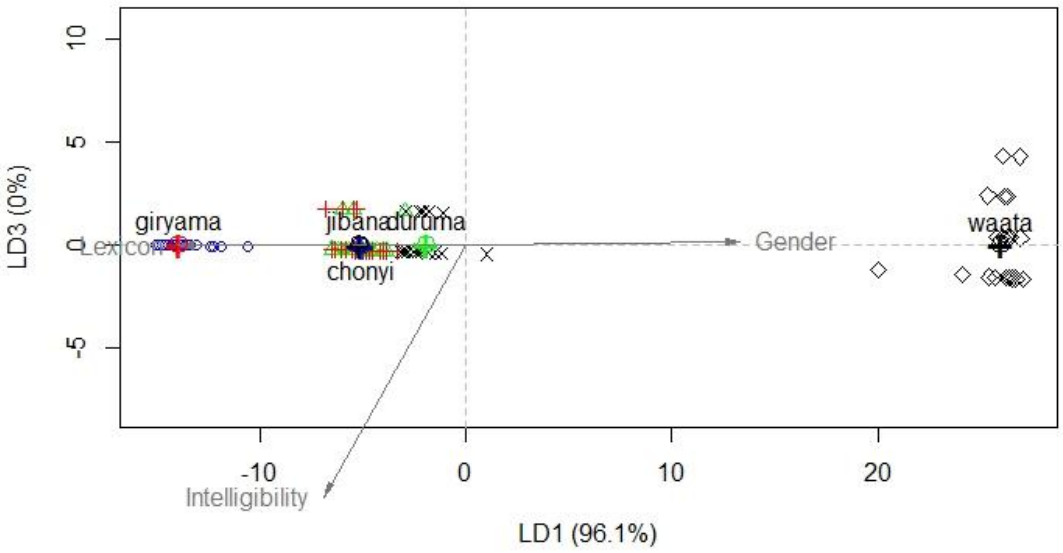


Figure 3: LD1 vs LD3

Figure 3 represents a plot of LD1 vs LD3 and it reveals some scatter within the groups. This is a clear indication that in this case the within-groups variances are not as minimised as in Figure 2.

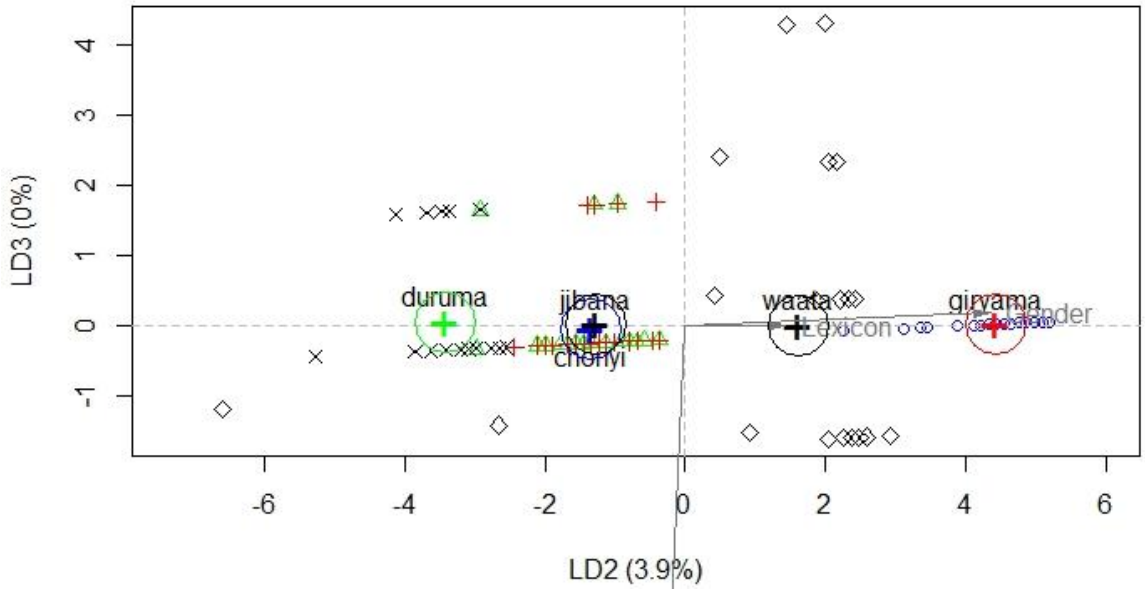


Figure4: LD2 vs LD3

A plot of LD2 vs LD3 as represented in Figure 4 above reveals that both the within-groups and the between-groups variances are not minimised. This is evident from the way the points are scattered, some of them mixing with points of other groups.

4. Discussion

4.1 Matrix scatterplot of raw data

There is some spread of points in cell two of row one in column two and cell one of row two in column one. The two cells are similar. More spread is observed in cell three of row two in column three, and cell two of row three in column two. However, there seems to be some grouping in cell one of row three in column one, and cell three of row one in column three.

4.2 Group means

The high Lexicon mean of 95.983333 for Giryama, and the relatively low Lexicon mean of 2.416667 for Waata could be an indicator that there is a distinction between the two groups based

on Lexicon. However, the other three language groups have their means so close in Lexicon such that it is hard to tell if they are distinct from each other or not.

Since Chonyi, Duruma, Giryama and Jibana return zero means on Gender while Waata returns a mean of 5.283333, it suggests that there could be a distinction between Waata and the other four language groups.

4.3 Contribution to group separation

In $Z_1$the absolute value of the coefficient for Gender (1.61409210) and the absolute value of the coefficient for Lexicon (0.33530441) are significant in the separation of the groups, whereas the absolute value of the coefficient for Intelligibility is almost zero, hence as per the MDA model, Intelligibility is not significant in the separation of the groups. This implies that Gender and Lexicon contribute immensely to the separation of the groups in that order, while Intelligibility does not contribute to the separation of the groups.

In Z_2,also the absolute value of Gender (3.34909990) and the absolute value of Lexicon (0.21786462) are significant whereas the absolute value of Intelligibility is almost zero. This implies that Gender and Lexicon contribute immensely in the separation of the groups in that order, whereas Intelligibility plays an insignificant role in the separation of the groups.

Z_3 is not important in the separation of the groups, as explained in *section 4.3* below, hence no need of discussing its coefficients.

4.4 Relative importance of each discriminant function

The proportions of traceimply that the first discriminant function, LD1, accounts for 96.11%, the second discriminant function, LD2, accounts for 3.89% and the third discriminant function, LD3, accounts for 0.00% of the separation of the groups. Thus, we deduce that LD1 and LD2best describe the separation of the groups while LD3does not.

4.5 The graphs of discriminant functions

As a result of applying MDA, separation leads to four distinct groups as observed in *graph 2*. Group one is composed of Giryama, group two is composed of Chonyi and Jibana, group three is composed of Duruma, while group four is composed of Waata. Chonyi and Jibana seem to be very close in characteristics hence they could not be distinguished as separate groups.

5 Conclusions

The univariate scatterplots of variables and the group of means suggest a distinction of the groups in Lexicon and Gender, but not in Intelligibility. Thus Intelligibility seems to play an insignificant role in the separation of the groups. However these techniques seem not to put a clear distinction between Chonyi, Jibana and Duruma because their means are close and the variables are assessed univariately.

The coefficients of the discriminant functionsshow that Lexicon and Gender play a significant role in the separation of the groups, whereas Intelligibility does not. Furthermore, a study of the plot of the discriminant functions shows that LD1 vs LD2 best minimises the within-group variances and maximises the between-groups variances. This results in four groups, that is, Giryama as one group, Chonyi and Jibana as one group, Duruma as one group and Waata as one group.We can therefore conclude that MDA can be used to distinguish between languages. Furthermore, the MDA model is precise, assesses all the three variables together, and thusbetter than the vastly used method of comparing means using of univariate techniques.

The findings of this research are in agreement with other findings by sociolinguists. On the one hand those sources reveal that Giryama, Chonyi, Jibana and Duruma are related and they belong to the larger Mijikenda group of languages. Furthermore, those sources indicate that Chonyi and Jibana are closely clustered together whereas Duruma and Giryama are distinct. All those four

languages are further classified as Bantu. On the other hand those sources classify Waata as a Cushitic language.(Gordon, 2018; Maho, 2009; Kenya Information Guide, 2015)

5.2 Recommendations

Based on the findings of this research, the following recommendations are hereby made to pave way for further research.

i. It is recommended that more research be conducted that shall further the application of MDA in other language groups to assess the performance. Furthermore, although this study was based on three variables, it is recommended that further research should involve more variables.

ii. This research was confined to the application of descriptive MDA, it is therefore recommended that research be conducted to apply predictive MDA in order to identify a sample with one of the known language groups.

Bibliography

[1]. Beine, D. (2017). *A social linguistic survey of the Bhatri-speaking communities of central India.* SIL International. Retrieved from https://www.sil.org/resources/publications/entry/70732

[2]. Blair, F. (1997). *Survey on a shoestring: A manual for small-scale language surveys.* Dallas: The Summer Institure of Linguistics and The University of Texas at Arlington.

[3]. Brown, C. E. (1998). *Applied Multivariate Statistics in Geohydrology and Related Sciences.* Springer.

[4]. Casad, E. (1974). *Dialect inteligibility testing.* Oklahoma: Summer Institute of Linguistics of the University of Oklahoma.

[5]. Faton, G. R. (2018). *Waci speakers in Togo and Benin: A sociolinguistic survey.* SIL International. Retrieved from https://www.sil.org/resources/publications/entry/74378

[6]. Gonzales, R. J. (2015). *Pasil survey report, Kalinga Province.* SIL International. Retrieved from https://www.sil.org/resources/publications/entry/69859

[7]. Gordon, R. G. (2018). *Ethnologue Languages of the World.* Dallas: SIL International. Imechukuliwa toka https://www.sil.ethnologue.com

[8]. Gries, S. T. (2015). The most under-used statistical method in corpus linguistics: multi-level (and mixed-effects) models. *Corpora, 10*(1), 95-125.

[9]. Jordan, L., & Manuel, I. (2016). *A sociolinguistic survey of Kwanza Sul Province, Angola: With a special focus on the Kimbundu variants.* SIL International. Retrieved from https://www.sil.org/resources/publications/entry/70691

[10]. Kassell, A., MacKenzie, B., & Potter, M. (2017). *Three Arafundi languages: A sociolinguistic profile of Andai, Nanubae and Tapei.* SIL International. Retrieved from https://www.sil.org/resources/publications/entry/74262

[11]. Kenya Information Guide. (2015). *The Mijikenda Tribe*. Imechukuliwa toka Kenya Information Guide: https://www.kenya-information-guide.com/mijikenda-tribe.html

[12]. Kluge, A. (2008). *Analysis of RTT results: Godirion survey example.* Chiang Mai: SIL International.

[13]. Kröger, O. (2005). *Report on a survey of Coastal Makua dialects.* Dallas: SIL International. Retrieved 2018, from https://www.sil.org/resources/publications/entry/9108

[14]. Lambrecht, P., & Mann, N. (2017). *Mel and Khaonh language survey report.* SIL International. Retrieved from https://www.sil.org/resources/publications/entry/70266

[15]. Lebold, R., & Lee, M. Y. (2016). *Survey report on the Emem language of Papua, Indonesia.* Dallas: SIL International. Retrieved 2018, from

https://www.sil.org/resources/publications/entry/70010

[16]. Maho, J. F. (2009). A classification of the Bantu languages: an update of Guthrie's referential system. Kwenye D. Nurse, & G. Philippson, *The Bantu languages* (kur. 639-651). London: Routledge.

[17]. Muniru, J., Magnusson, C., Hansley, M., & Ayenajeh, S. (2016). *A sociolinguistic survey of the Kulere dialects of Plateau and Nassarawa states, Nigeria.* Dallas: SIL International. Retrieved 2018, from https://www.sil.org/resources/publications/entry/69738

[18]. Rencher, A. C. (2002). *Methods of multivariate analysis* (2nd ed.). New York: John Wiley & Sons.

[19]. *Statistics Solutions*. (2018). Retrieved May 14, 2018, from Statistics Solutions: http://www.statisticssolutions.com/discriminant-analysis/

[20]. Webster, J. (2017). *A sociolinguistic profile of the Tharu dialects of the Western Indo-Nepal Tarai.* SIL International. Retrieved from https://www.sil.org/resources/publications/entry/70672