Clusters of Leading Death Causes in South Africa: Application of Hierarchical Agglomerative Clustering Technique

Ntebogang Dinah Moroke (Dr)

Malebogo Pulenyane

Pulenyane Malebogo, North West University, RSA, Private Bag X2046 Mmabatho, 2735
Email: Ntebo.Moroke@nwu.ac.za

Doi:10.5901/mjss.2014.v5n20p848

Abstract

This paper presents an exploratory method for investigating the structure underlying the data. The methods used are reported effective for finding similarity between groups of cases or variables. Furthermore, these methods (hierarchical agglomerative clustering algorithm) are useful when a priori groups are unknown. The results from these methods are in a form of clusters presented in a hierarchy-like structure. Data consisting of 537 out of 1079 variables collected from January to December in 2009 by the Department of Home Affairs, disseminated by Statistics South Africa head office was analysed using SPSS 22. A dendogram of a single linkage method from the hierarchical agglomerative algorithm revealed the five clusters formed from the 537 leading death causes. These causes were collected in clusters according to their hazards with respiratory tuberculosis and pneumonia as main leading causes of death followed by diarrhea, stroke and heart failure. The clusters formed were validated using discriminant analysis which reported about 0.4% of classification error rate. Wilk's Lambda proved that all the clusters were significant accordingly. While long term plans can be secured for death causes in the fifth cluster, it is important to pay special attention to death causes in clusters 1 to 4 urgently, more specifically those in the first cluster. This may reduce death rates in the country and life spans of residents may also be prolonged. Further analysis may be done where these clusters will be used as variables. A confirmatory factor analysis may be used to further confirm these clusters.

Keywords: *Clusters, causes of death, dendogram, hierarchical agglomerative clustering, South Africa*

1. Introduction

Cluster analysis is one of the multivariate exploratory techniques used for discovering groups of similar observations within a data set. Clustering is done to form groupings of objects in such a way that similar objects will belong to the same group than those in the other group. Apart from identifying similarities and developing subgroupings of homogenous entities, cluster analysis may also be worthwhile in finding the true groups that are assumed to truly exist and may also be useful for data. This involves determining differences and similarities among variables in multivariate space. Areas of study that are faced with large sets of data may find this tool very useful. Inconsistencies and complexities in the data may be addressed when this tool is used as a precursor in data analysis.

Owing to the inherent complexity in multivariate data, it is often desirable to find relationships among a suite of variables from which patterns or structures can be determined. This may be done either to gain a more thorough understanding of outcome variables or to develop groups that can be subjected to further analyses (Cross, 2013).

This paper applies the hierarchical agglomerative clustering method to what constitute leading death causes in South Africa (SA). Although SA has a functioning death registration system, the quality of cause of death data has been questioned. This is due to data from demographic surveillance studies using verbal autopsies to determine cause-specific mortality according to Adjuik *et al*. (2006). The available system does not indicate or identify the more deadly death causes or does not show groupings of these causes accordingly. World Health Organization (WHO)[1] is constantly monitoring improvements of data on causes of death. Some of the causes may be similar but the ICD-10 coding system is unable to identify them. As a result, this study uses cluster analysis (CA) to investigate these causes and attempts to identify the similarities and differences between the diseases. CA help in eliminating the duplication of the recordings and may also make life easier for doctors and other responsible authorities when finding cure for such causes. The findings of

[1] *The United Nations agency coordinates international health activities and helps governments improve health services. This agency checks the quality of the identification of causes of death and the coding system used in a particular country.*

this study may help this organisation to get a clear picture about the groupings of death causes in the country. Better approaches to prevent or reduce high rates of death in the country may be developed. This will help responsible personnel when making short and long term plans. An urgent attention may be paid to the most malicious causes.

CA has been applied in several areas such as computer science by Wallace, Keil & Rai, (2004), neuropsychology (Allen, Goldstein & Warnick, 2003; Allen *et al*. 2010; Thaler *et al*. 2010), biology (Eisen, Spellman, Brown & Botstein, 1998; Jiang, Tang & Zhang, 2004). These sources may be consulted for further advancement of knowledge on the subject.

The rest of this paper is organised as follows: Section 2 briefly discusses the hierarchical clustering technique and data used. Section 3 presents and discusses the results while section 4 gives concluding remarks.

2. Data

The data used in this study is records on mortality and causes of death in SA collected by the Department of Home affairs. After verifying and validating the data using the framework proposed by Mahapatra *et al*. (2007), Statistics South Africa (Stats SA) head office published the data with compact disc and has it available to users on request. The original list consists of 1079 mortality and death causes which took about 572 673 lives in the country. The data was recorded from January to December of the year 2009 by age, sex, population group, marital status, place or institution of death occurrence, province of death occurrence and province of usual residence of the deceased (Stats SA release, 2010). After filtering was done by removing all those causes that has as few as about 5 cases, 537 (about 50% of the causes) were left to be used in the analysis. The data is captured and analysed using Statistical Packaging for Social Sciences (SPSS) version 22.

One of the assumptions about data used for CA is the absence of outliers. Though outliers have a tendency of influencing the clusters, variables and/or cases containing outliers will not be removed. In cluster analysis, an outlier can describe a case that is either an extreme value within its own cluster or a value so extreme as not to belong to any cluster (Leonard and Droege, 2008). It is expected that some variables have large values meaning that most deaths were as a result of that particular death cause. Likewise, causes with low values imply that very few people have died as a result of that death cause. Instead of removing cases or variables that constitute outlier(s), the data will be standardised using z-scores. This transformation method is also chosen since all the variables used in this analysis are measured on a continuous scale. There are no missing values in the data. All blank spaces imply that there is no death recorded for that variable or people did not die of such a cause during that particular year.

3. Theoretical Framework

This section gives a brief review of the theoretical framework used in the analysis.

3.1 Hierarchical cluster method

CA is used to uncover pattern in the data. This technique is also used to reveal associations between different variables. Everitt (2010) suggested CA as a method for uncovering groups or clusters of observations that are homogeneous. The clusters are more visible when represented in a form of hierarchy. Johnson (1998) defines hierarchical clustering as observed data points grouped into clusters in a nested sequence of clustering. These methods reveal what looks like a hierarchical tree-like structure (Lattin *et al.* 2003). For more defintitions and extensions of the definition by Johnson can be found in CAbooks by Everitt, Landau, Leese & Stahl (2011).

There are two types of Hierarchical clustering procedures; the agglomerative method which merges clusters according to the distances between them and the divisive method which splits the already formed clusters. Adaekalavan & Chandrasekar (2013) suggested the Hierarchical Agglomerative Clustering as a method of connectivity. This method has been found efficient in computation and discovering the clusters in a set of data. Owing to the non-feasibility of examining all possible clustering possibility for a large set of data, Rencher and Christensen (2012) showed that the number of ways of dividing the (*n*) objects into *g* clusters can be illustrated as:

$$N(n,g)=\frac{1}{g!}\sum_{k=1}^{g}\binom{g}{k}(-1)^{g-k}k^{n}, \qquad (1)$$

where $k=1,2,.....,m,$ n is the number of observations, and g is number of clusters formed. Rencher and Christensen explain that (1) can be approximated by $g^n/g!$ which is large for even moderate values of *n* and *g*. As a

result, hierarchical methods permits allow determining the reasonable without having to look at all possible arrangements.

As explained, hierarchical methods involve a sequential process whereby an observation or a cluster of observations is merged into another cluster in each step of agglomerative hierarchical approach. As the process unfolds, the number of clusters reduces and the clusters themselves grow larger. The process starts with n clusters that constitute the individual items and remain with one cluster containing all the items in a set of data. The initial step of agglomerative hierarchical procedure merges the clusters according to the minimum distance to form a new cluster. This process is irreversible in such a way that items that are lumped together to form a cluster cannot be separated later in the procedure (Hair *et al.* 2010, Rencher, 2002 and Rencher and Christensen, 2012).

The similarity or dissimilarity of two clusters is measured using a single linkage method in this study. This method merges the items according to a minimum distance that separates them. For example, the distance between points A and B as illustrated by Rencher and Christensen can be shown as:

$$D(A,B) = \min\{d(y_i, y_j), \text{for } y_i \text{ in } A \text{ and } y_j \text{ in } B, \quad (2)$$

where $d(y_i, y_j)$ is the Euclidean distance calculated as:

$$d(x,y) = \sqrt{\sum_{i=1}^{p}(x_j - y_j)^2} \quad (3)$$

At each step, the distance (2) is computed for every pair of clusters and the merging of two clusters and done with the smallest distance. The number of clusters is reduced by one at each step and the process is repeated for the next step. The process is stopped if number of clusters equals one (Everitt, 2010).

3.2 Interpretation of clusters

Interpretation involves examining each cluster in terms of the cluster variate to identify the characteristics of each cluster. This also helps in describing the nature of clusters (Johnson and Wichern, 2007). Asheim, Cooke and Martin (2006) cautioned that it is not always easy to construct a single theory for all the objects in a cluster.

3.3 Presentation of results

The results of hierarchical clustering process can be displayed graphically in a tree diagram also known as a dendogram. According to Sneath and Sokal (1973) and Hartigan (1975), this diagram allows all the steps in the hierarchical procedure including the distances at which clusters are merged. To be precise, a dendogram is constructed from $n \times n$ distance matrix and it illustrates how agglomeration takes place (Bryan, 2005). Firstly, items are arranged hierarchically such that those with the highest mutual similarity are placed together. The next step collects objects that are more closely associated with those in other groups. This procedure continues until all the individuals have been placed into a complete classification scheme (Isah, Abdullahi and Waziri, 2013). The degree of similarity is depicted in the dendogram by length of branch.

3.4 Validity if clusters

Discriminant analysis used in this study to check the convergent and classification validity of death causes to identified clusters. This method helps in confirming if variables converge together as expected (convergent validity) and as reported by the analysis. Furthermore, this statistic provides a basis for verifying the statistical significance of each cluster. This statistic ranges between zero and one, with a value closer to zero denoting a high level of significance of that cluster. The observed values are compared with critical values using a significance level of 5%.

The following section presents the results from the hierarchical clustering. The results are summarised in tables.

4. Empirical Results

This section provides selected results from the analysis. Prior the analysis, a query was submitted to SPSS to standardize the variables as discussed. The standardized data was used to obtain the results from the hierarchical aagglomerative clustering and the results are presented in Table 1 though Table 3. *Note that the tables are pasted in this document as a picture due to their size*. Another preliminary analysis of data involves presentation of the descriptive

statistics[2] of the 537 variables used in the analysis. This gives a picture of the distribution of the variables and helps in making their descriptive comparison.

Observing the output on descriptive statistics (not shown in this paper) shows that for each of the 537 variables, there are 12 valid cases (number of months for which death was recorded). Also revealed is the fact that respiratory tuberculosis, diarrhea and gastroenteritis of presumed infectectious origin are the most leading causes of death in SA. The mean and variances of these variables are higher than those of other variables. These findings are not different from those by Groenewald *et al*. (2010) who reported respiratory tuberculosis to be the third leading cause of death in the Cape Town Metro district in 2006. Same findings were reported by Bradshaw, *et al*. (2003) on burden of disease estimates for SA with focus on burdens due to premature mortality. Other causes that lead to many families loosing their loved ones in 2009 are heart failure, stroke and respiratory failure. All these causes have averages in the thousands. Some of the least threatening causes of death are among others hypothermia of newborn, malformations of lung, diverticular disease of intestine, etc. When some of the variables are defined by a normal distribution, few reveal a left skewness and others a right skewness. This simply means that that as much as there are similarities in the causes of death, there are also causes with some differences, a good indication that the use of CA to this data is a good idea.

Table 1 presents the results from the hierarchical agglomerative clustering algorithm. Firstly, the Euclidean distances were calculated for pairs of clusters according to (3). This matrix could not be shown here due to its size. This matrix corresponds to the number of death causes being clustered (537 by 537 matrix). The results of this matrix reported variety of distances ranging between 2 and 7557.869. The matrix is used as a starting point for clustering. Clusters between the variables are formed according to the smallest distance separating them. The distances can clearly be seen in Table 2 column 3.

Table 2: Agglomeration schedule

Stage	Cluster Combined		Coefficients	Stage Cluster First Appears		Next Stage
	Cluster 1	Cluster 2		Cluster 1	Cluster 2	
1	134	292	2.000	0	0	26
2	2	44	2.236	0	0	34
3	376	521	2.236	0	0	11
4	359	516	2.236	0	0	8
5	191	363	2.236	0	0	8
6	334	348	2.236	0	0	13
7	298	340	2.236	0	0	10
8	191	334	2.236	5	6	19
9	166	304	2.236	0	0	17
10	68	191	2.236	9	0	12
11	3	359	2.449	0	4	15
12	3	350	2.449	11	0	27
13	16	298	2.449	0	7	15
14	256	262	2.449	0	0	20
15	3	256	2.449	12	14	17
16	134	169	2.449	1	0	19
17	68	134	2.646	16	0	33
18	382	532	2.646	0	0	27
19	68	524	2.646	17	0	23
20	3	442	2.646	15	0	33
21	366	378	2.646	0	0	111
22	193	282	2.646	0	0	24
23	3	252	2.646	20	0	26
24	3	242	2.646	23	0	61
25	97	233	2.646	0	0	40
26	2	122	2.646	2	0	28
27	16	68	2.646	13	19	30
28	3	16	2.828	0	27	51
29	176	525	2.828	0	0	31
30	3	458	2.828	28	0	32
31	3	457	2.828	30	0	34
32	3	436	2.828	31	0	36
33	366	422	2.828	21	18	36
34	3	376	2.828	32	0	39
35	185	3	2.828	0	0	37
36	3	366	2.828	34	33	40
37	317	371	2.828	0	0	49
38	132	297	2.828	0	0	54
39	42	185	2.828	36	0	46
40	2	3	2.828	26	0	102
41	323	449	3.000	0	0	110
42	239	435	3.000	0	0	50
43	199	415	3.000	0	0	66
44	323	373	3.000	0	0	62
45	215	342	3.000	0	0	48
46	2	332	3.000	0	0	117
47	71	300	3.000	0	0	49
48	2	296	3.000	0	0	69
49	2	289	3.000	0	0	51
50	2	199	3.000	0	43	52
51	2	176	3.000	0	29	53
52	2	131	3.000	0	0	54
53	2	124	3.000	0	0	56
54	2	42	3.162	0	39	61
55	439	517	3.162	0	0	59
56	2	501	3.162	54	0	73
57	26	498	3.162	0	0	63
58	2	492	3.162	56	0	124
59	341	440	3.162	0	0	62
60	377	438	3.162	0	0	69
61	149	377	3.162	0	55	96
62	215	374	3.162	0	60	64
63	2	372	3.162	58	0	65
64	2	344	3.162	63	0	66
65	2	323	3.162	64	44	67
66	2	305	3.162	65	0	68
67	2	170	3.162	67	0	71
68	110	149	3.162	0	0	70
69	77	110	3.162	0	61	103
70	2	103	3.162	0	69	72
71	2	68	3.162	0	0	73
72	2	56	3.162	71	0	74
73	2	34	3.162	72	57	76
74	2	13	3.162	73	0	98
75	162	515	3.317	0	0	77
76	2	464	3.317	74	0	80
77	2	385	3.317	76	0	81
78	306	380	3.317	0	0	82
79	115	351	3.317	0	0	95
80	2	326	3.317	77	0	105
81	97	306	3.317	78	0	106
82	115	295	3.317	79	0	88
83	151	272	3.317	0	0	141
84	121	247	3.317	0	80	86
85	2	181	3.317	0	0	97

NB: Only information on the first 85 stages of Hierarchical agglomerative clustering was shown. This is due to the size of the table.

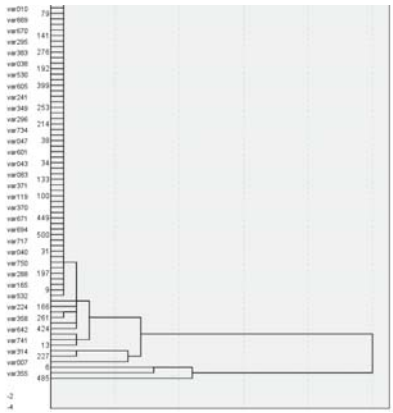
In Table 1 at Stage 1, Variable 134[3] is clustered with Variable 292. The minimum Euclidean distance between these two variables is 2. Neither variable has been previously clustered as seen in column 5 and 6, the two zeros under Cluster 1 and Cluster 2. The next stage when the cluster containing variable 134 combines with variable 169 by a distance 159 is Stage 16. Variable 134 merges with variable 68 at stage 17 and the latter appears next at stage 19 forming a new cluster with variable 524 through a minimum distance 2.646.

The largest cluster is a combination of variable 1 and 6 with a distance 7557.869. Variable 1 made first appearance at stage 314 with variable 2 through a distance 11.225 and variable 6 is making first appearance at the last stage 537. No further analysis is possible after stage 537.

A dendogram shown as Figure 1 is a graphical representation of agglomerative schedule, Table 1. Number of clusters is visible in this diagram. A dendogram, due to its branching-type nature, allows one to trace backward or forward to any individual case or cluster at any level. Additionally, it gives an idea about the magnitude of the distance between

[2] *See appendix for selected descriptive statistics*
[3] *See appendix for selected variable numbers.*

cases or groups that are clustered in a particular step, using a 0 to 25 scale along the top of the chart. While it is difficult to interpret distances in the early clustering phases shown on the extreme left of the plot, relative distances become more apparent as you move to the right. The bigger the distances before two clusters are joined, the bigger the differences in these clusters. A membership of a particular cluster can simply be traced backwards down the branches to the name. As seen in the dendogram, five clusters are visible, implying that leading death causes in SA can be classified into three groups.



NB: Due to large size, the diagram was cropped.

Figure 1: Dendogram of a Single Linkage Method

Presented in Table 2 and 3 is the output used to assess the validity of the five clusters of leading death causes in SA.

Table 2: Wilks' Lambda

Test of clusters (s)	Wilks' Lambda	Chi-square	df	Sig.
1 through 4	.001	3448.420	48	.000
2 through 4	.026	1916.343	33	.000
3 through 4	.152	994.488	20	.000
4	.588	280.044	9	.000

Though the variables were collected in five clusters, SPSS reveals the results of only four clusters. These results reveal that all the four clusters are significant at all levels of significance according to the observed probabilities associated with Wilk's Lambda. This is confirmed by the canonical correlation coefficients in Table 3 showing the first three clusters with high correlations and the last one with a moderate correlation. The variables in cluster one explains about 68% of variation to leading causes of death with cluster four showing the least contribution of not more than 1%. Observing the classification status of death causes to clusters confirmed an apparent error rate of 0.06%, implying that most of these variables are correctly classified (correct classification rate of 99.6%) as members of respective clusters. One of the variables was incorrectly classified in cluster number 2 hence the results.

Table 3: Eigenvalues

Cluster	Eigenvalue	% of Variance	Cumulative %	Canonical Correlation
1	17.254[a]	67.5	67.5	.972
2	4.741[a]	18.5	86.0	.909
3	2.874[a]	11.2	97.3	.861
4	.700[a]	2.7	100.0	.642

a. First 4 canonical discriminant functions were used in the analysis.

5. Concluding Remarks

This paper assessed the effectiveness of hierarchical agglomerative clustering technique on the 5.37 leading causes of death in South Africa for the year 2009. A sample of about 50% was reached after filtering was done to the 1079 death causes. Those causes that lead to as few as two deaths were not included in the analysis. Due to the large dataset

analysed, some of the tables could not be shown in this paper and some were pasted using an enhanced metafile. A dendogram revealed about five clusters explaining the 537 leading death causes. The clusters collected the causes according to their hazard with the first cluster showing most dangerous and the last containing the least dangerous causes.

The results may be used by policy makers in the Department of Health to embark on policies that may prevent death causes especially those in Clusters one to three. This may mean more qualified personnel to be employed in health care centres in the country and those present to be equipped with necessary skills. Building of more specialised health care centres may also help reduce the risks as people will be able to access them quickly and with ease. Serious awareness should be made to residents about these causes as they affected many families and the trends seem to be increasing every year.

For further studies, other multivariate techniques such as multiple regression and discriminant analyses may be used where clusters will be used as independent variables. Structural equation modelling techniques may also be used to confirm the results of this study.

6. Scope Limitations

The results discussed in this document were based on information on mortality and causes of death in SA obtained from the civil registration system. The main focus was on deaths that were officially recorded in 2009 and released during the 2010/11 processing phase. Though stillbirths were also recorded during that period, the allied statistics are excluded in the analysis.

7. Acknowledgements

The authors are grateful to the Department of Home Affairs for collecting the data and Statistics South Africa for disseminating the dataset used in this study.

References

Adaekalavan, S. & Chandrasekar, C. (2013). Towards new estimating in cremental dimensional algorithm (EIDA). *Journal of Theoretical and Applied Information Technology*, 51 (2): 222-227.

Adjuik, M., Smith, T., Clark, S., Todd, J., Garrib, A., Kinfu, Y., Kahn, K., Mola, M., Ashraf, A., Masanja, H., Adazu, U., Sacarlal, J., Alam, N., Marra, A., Gbangou, A., Mwageni, E. & Binka, F. **(**2006). Cause-specific mortality rates in sub-Saharan Africa and Bangladesh. *Bulletin of the World Health Organization,* 84: 181-188.

Allen, D. N., Goldstein, G., & Warnick, E. (2003). A consideration of neuropsychologically normal schizophrenia. *Journal of the International Neuropsychological Society, 9*: 56–63.

Allen, D. N., Leany, B. D., Thaler, N. S., Cross, C., Sutton, G. P., & Mayfield, J. (2010). Memory and attention profiles in pediatric traumatic brain injury. *Archives of Clinical Neuropsychology, 25*: 618–633.

Asheim, B., Cooke, P. & Martin, R. (2006). Clusters and Regional Development: Critical reflections and explorations. Routledge, United States of America and Canada.

Bradshaw, D., Groenewald, P., Laubscher, R., Nannan, N., Nojilana, B., Norman, R., Pieterse, D. & Schneider, M. (2003). Initial burden of disease estimates for South Africa, 2000. MRC Technical Report. Cape Town: Medical Research Council.

Bryan, F.J.M. (2005). *Multivariate Statistical Methods*. A Primer (3rded).Chapman and Hall/CRC, USA.

Cross, L.C. (2013). *Statistical and Methodological Considerations When Using Cluster Analysis in Neuropsychological Research*. Springer Science and Business Media. New York. 978-1-4614-6744-1.

Eisen, M.B., Spellman, P.T., Brown, P.O. & Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences, 95*: 14863– 14868.

Everitt, B.S. (2010). *Multivariate modelling and multivariate analysis for the behavioural sciences*, Taylor and Francis Group. LLC, United States of America.

Everitt, B.S., Landau, S., Leese, M., & Stahl, D. (2011). *Cluster analysis* (5th ed.). New York: Wiley.

Groenewald, P., Bradshaw, D., Daniels, J., Zinyakatira, N., Matzopoulos, R., Bourne, D., Shaikh, N. & Naledi, T. (2010). Local-level mortality surveillance in resource-limited setting: a case study of Cape Town highlights disparities in health. *Bull World Health Organ*, 88: 444- 451.

Hair, J.F., Black, W.C., Babin, B.J., Anderson, R.E. (2010). *Multivariate data analysis*: A global perspective (7thed). Pearson Prentice Hall, Upper Saddle River.

Hartigan, J. A. (1975). *Clustering Algorithms*. New York: Wiley.

Isah, A., Abdullahi, U. & Waziri, V.O. (2013). A hierarchical cluster analysis and simulation of state capitals in Nigeria for tourism exploration. *International Journal of Latest Research in Science and Technology*, 2(1), 437-441.

Jiang, D., Tang, C., & Zhang, A. (2004). Cluster analysis for gene expression data: A survey. *IEEE Transactions on Knowledge and Data Engineering, 16*: 1370–1386.

Johnson, D.E. (1998). *Applied multivariate methods for data analysis*. Brooks/Cole publishing: United States of America.

Lattin, J., Carroll, J.D. & Green, P.E. (2003). *Analyzing multivariate data*. Brooks/Cole publishing. Canada.

Leonard, S.T. & Droege, M. (2008). The uses and benefits of cluster analysis in pharmacy research. *Research in Social and Administrative Pharmacy*, 4: 1-11.

Mahapatra, P., Kenji, S., Alan, D.L., Francesca, C., Francis, C.N. Simon, S. on behalf of the Monitoring Vital Events (MoVE) writing group (2007). Civil registration systems and vital statistics: successes and missed opportunities. *The Lancet*, 370 (10): 1653-1663.

Rencher, A.C., & Christensen,W.F.(2012). *Methods of Multivariate Analysis*, (3rded). Wiley-Interscience, ISBN 9780470178696.

Rencher, A.C. (2002). *Methods of Multivariate Analysis.* John Wiley & Sons, Inc: Canada.

Sneath, P. H. A. & Sokal R.R. (1973). *Numerical Taxonomy*. San Francisco: Freeman.

Statistics South Africa (2010/11). Mortality and causes of death in South Africa, 2009: Findings from death notification. Statistical release P0309.3. Pretoria: Statistics South Africa.

Thaler, N. S., Bellow, D. T., Randall, C., Goldstein, G., Mayfield, J., & Allen, D. N. (2010). IQ profiles are associated with differences in behavioral functioning following pediatric traumatic brain injury. *Archives of Clinical Neuropsychology, 25*: 781–790.

Wallace, L., Keil, M., & Rai, A. (2004). Understanding software project risk: A cluster analysis. *Information and Management, 42:* 115–125.

Appendix: First 41 causes of death in SA

No.		Mean	Std. Deviation
1	Other ill-defined and unspecified causes of mortality (R99)	5509.5	656.76653
2	Pneumonia; organism unspecified (J18)	4586.4167	647.06238
3	Respiratory tuberculosis; not confirmed bacteriologically or histologically (A16)	4120.5	222.01658
4	Stroke; not specified as haemorrhage or infarction (I64)	1944.6667	204.12934
5	Heart failure (I50)	1841.25	238.59594
6	Diarrhoea and gastroenteritis of presumed infectious origin (A09)	1809.3333	369.36883
7	Exposure to unspecified factor (X59)	1312.0833	107.35113
8	Cardiac arrest (I46)	1299.9167	110.34612
9	Other septicaemia (A41)	1158.5	38.69578
10	Respiratory failure; not elsewhere classified (J96)	986.3333	97.13377
11	Stillborn (P95)	961.9167	66.65714
12	Acute myocardial infarction (I21)	869.5	111.93302
13	Unspecified acute lower respiratory infection (J22)	826.4167	101.09623
14	Meningitis due to other and unspecified causes (G03)	693.75	52.89118
15	Volume depletion (E86)	610.4167	87.68586
16	Unspecified diabetes mellitus (E14)	585.5833	70.28961
17	Hypertensive heart disease (I11)	569.1667	93.15464
18	Essential (primary) hypertension (I10)	542.75	72.9497
19	Unspecified renal failure (N19)	507.3333	23.20397
20	Miliary tuberculosis (A19)	461.75	49.98204
21	Other viral diseases; not elsewhere classified (B33)	395.0833	31.82469
22	Discharge from other and unspecified firearms (W34)	388.5833	20.96082
23	Motor- or nonmotor-vehicle accident; type of vehicle unspecified (V89)	372.5	51.87835
24	Other respiratory disorders (J98)	362	57.25065
25	Contact with blunt object; undetermined intent (Y29)	357.5	36.15497
26	QCL_2	353.9889	220.73335
27	Assault by sharp object (X99)	351.75	70.65939
28	Other symptoms and signs involving the circulatory and respiratory systems (R09)	326.9167	37.68158
29	Malignant neoplasm without specification of site (C80)	318.25	28.45131
30	Tuberculosis of nervous system (A17)	310.25	19.30203
31	Other chronic obstructive pulmonary disease (J44)	306.9167	61.74207
32	Asthma (J45)	303.25	35.75453
33	Other accidental hanging and strangulation (W76)	298.0833	41.2056
34	Other disorders of fluid; electrolyte and acid-base balance (E87)	296.6667	29.05898
35	Other immunodeficiencies (D84)	291.5	26.27304
36	Malignant neoplasm of bronchus and lung (C34)	288.0833	19.31772
37	Other anaemias (D64)	264.3333	26.13369
38	Pneumocystosis (B59)	257.8333	16.13485
39	Disorders related to short gestation and low birth weight; not elsewhere classified (P07)	255.5833	33.18121
40	Hepatic failure; not elsewhere classified (K72)	253.75	15.02195