Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Modeling Mortality by Causes of Death in South Africa Using Log-linear Analysis

Exaverio Chireshe (46520996@mylife.unisa.ac.za)
University of South Africa

Research Article

Keywords: causes of death, mortality modeling, log-linear analysis, backward elimination, forward selection, odds ratios, contingency table, goodness of fit test, parameter estimation, Pearson chi-square, likelihood ratio, Akaike's information criterion, Bayesian's information criterion, residual deviance.

Posted Date: February 15th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-243579/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

MODELING MORTALITY BY CAUSES OF DEATH IN SOUTH AFRICA USING LOG-LINEAR ANALYSIS

by

EXAVERIO CHIRESHE:

UNIVERSITY OF SOUTH AFRICA

FEBRUARY 2021

Abstract

This study aimed at identifying underlying patterns in mortality due to causes of death in South Africa using mortality statistics from 2005 to 2015 obtained from Statistics South Africa. Log-linear analysis was used in this study on mortality by causes of death dataset having three variables, cause of death (C), province (P) and year (Y). Log-linear analysis was preferred because of its capability to tease out relationships among variables. Results revealed that there are variations in mortality due to causes of death. Mortality was found to differ widely across the country, among provinces. It is recommended that prevention and management policies for HIV and TB be intensified since they still remain South Africa's major causes of death. A replication of the study could be done in another developing country using latest data to see if it will yield the same results. A multi-population mortality modeling could also be carried out using the same approach.

Keywords: $causes\ of\ death$, $mortality\ modeling$, $log-linear\ analysis$, $backward\ elimination$, $forward\ selection$, $odds\ ratios$, $contingency\ table$, $goodness\ of\ fit\ test$, $parameter\ estimation$, $Pearson\ chi-square$, $likelihood\ ratio$, $Akaike's\ informartion\ criterion$, $Bayesian's\ information\ criterion$, $residual\ deviance$.

1. Introduction

The ability to survive for a long time having a healthy life is an important element of human betterment. The 2^{nd} part of the 20^{th} century saw huge progress in upgrading health and life span across the globe. The United Nations (UN) (2014) noted that the world population's longevity at birth rose to 68 years in 2005 - 2010 from 48 years in 1950 - 1955. According to Alicandro et al. (2018), there are comprehensive differences in mortality levels amongst nations and regions. The differences may be due to disparities in access to potable water, medical care, food and sanitation. These differences may also be a result of risk factors, societal issues and behavioral choices which influence individuals' survival (Alicandro et al., 2018).

Chavhan and Shinde (2016) pointed out that forecasting and modeling mortality have been a major study area for the demographers and actuaries. According to these authors, a lot of states or nations have witnessed significant improvements in death rates during the last century, resulting in improved life expectancy.

Ritchie and Roser (2018) argued that there has been changes in causes of death across the world in the past few years due to improved standards of living and as a result the life expectancy scaled up. Furthermore, the authors noted that out of all the global deaths for 2016, non-communicable diseases also known as chronic diseases dominated for both mortality numbers at global level and also deaths in developed nations. Amongst underdeveloped and developing countries, some of the causes of death are still common and in some cases still dominant. In South Africa, HIV and TB infections remain the major causes of death (Phetlhu et al., 2018).

In the public health sector, many researches have been carried out to analyse mortality by causes of death. Availability of causes of mortality and mortality statistics has enabled researchers to come out with speculative models to interpret mortality trends. Most of the models constructed were mainly about multivariate statistics such as linear regression,ordinary least squares regression (OLSR), analysis of variance (ANOVA) and many others. Recent models include logistic model, multilevel model and the Bayesian hierarchical model. Yifang (2013) used both logistic and multilevel models to analyse child mortality in Nigeria. Li et al. (2019) proposed a forecast reconciliation model to cause of death mortality modeling using the USA population. Alicandro et al. (2018) used Poisson regression models to detect major causes of death contributing to absolute and relative socio-economic inequality in Italy. Li, O'Hare and Zhang (2015) applied a semiparametric varying panel model to mortality modeling using mortality statistics of various first world countries. Karimi, Rey and Latouche (2017) proposed a joint modeling of socio-professional and cause-specific mortality using the French population.

Among various methods for the quantitative study of data that is categorical, log-linear models currently play a crucial role in social statistics, their sophistication and complexity having swiftly developed over the past thirty years (Manzo, 2014). Log-linear models are debatably the most famous and essential statistical models for analysing categorical data. It is an advanced technique to determine elements that subtend the relative frequency of several attributes. The application of log-linear modeling has been proposed as a statistical analysis technique when dealing with categorical count data (Lacobucci & McGill, 1990). Olmus and Erbas (2012) argued that log-linear modeling has an added advantage over other techniques because of its powerful statistical weight. A distinctive attribute of log-linear modeling is its capability to tease out associations across variables' responses and the factorial arrangement of the study design categories. Log-linear models explains interrelationships and association trends across a set of categorical factors.

Nyman et al. (2015) argued that log-linear models are the famous pillars for analysing multi-way tables. According to Nyman et al. (2015), a log-linear parameterisation of an association model can be more revealing than just a direct parameterisation on the basis of likelihoods, resulting in an effective way of explaining restraints inferred from conditional, marginal and context-specific

independence. The application of log-linear analysis in mortality and causes of death is a recent development. Adarabioyo (2014) applied log-linear modeling to determinants of child mortality in Nigeria.

Various approaches have been employed to model mortality in South Africa. Johnson et al. (2017) used Bayesian approach in South Africa to evaluate the effect of antiretroviral treatment on mortality patterns for adults. A study by Rademeyer (2017) on provincial differentials in under five mortality in South Africa used negative binomial regression. Oldenburg et al. (2018) carried out a research on mortality and antiretroviral therapy in South African rural areas using inverse probability weighted (IPW) marginal structural models (MSMs) approach and regression discontinuity design (RDD) approach. Manda and Abdelatif (2017) applied a hierarchical Bayesian shared component spatial-temporal model using the municipality as the spatial unit for analysis on the study of smoothed temporal atlases of age-gender all cause mortality in South Africa. In another study, Chikobvu and Shoko (2018) used principal component analysis approach to predict mortality caused by HIV and AIDS applying CD4 cell counts based states and viral load in South Africa. A research by Pillay-van Wyk et al. (2016) applied multinomial logistic regression model to model the 2009 injury mortality survey data by age, sex, province and population group.

Musenge et al. (2012) applied Integrated Nested Laplace Approximation (INLA) and Stochastic Partial Differential Equations (SPDE) approaches to model HIV/TB child mortality from 1992 to 2010 in South Africa. Udjo and Lalthapersad-Pillay (2013) used Growth Balance Method and Gompertez model to estimate maternal mortality and causes in South Africa. In another study, Scovronick et al. (2018) used time series analysis to determine the association between ambient temperature and mortality in South Africa.

Out of all the researches carried out to model mortality in South Africa, not even a single research employed log-linear approach. The study of mortality and causes of death using log-linear modeling will compliment the existing knowledge on the subject in a South African context. The study can also be utilised to craft curative and preventive methods or determine resource allocation aimed at increased life expectancy. The analysis may also be used to draw conclusions about future mortality trends in South Africa.

2. Methodology

2.1 Introduction

Log-linear modeling was employed to test the influence of variables: province, year and cause of death on mortality. Materials and approaches applied to present and analysing the data using log-linear analysis are described following the 6 stages in building a model suggested by Hair et al. (2019). These authors proposed six steps in building a model which include research objectives, design of the research, testing of assumptions, model estimation and fitting, validation of findings and lastly analysis and results discussion. Focus of this section is on outlying research objectives, defining the design of the research, estimating the log-linear models and assessing the overall fitness of the model.

2.2 Research objectives

Log-linear modeling was used in this study with the aim of obtaining a less complex model that adequately accounts for the data. Model elements which are essential to be incorporated in the model are easily identified through employing log-linear method.

2.2.1 Objectives

The major objectives were to:

1. Find out whether or not there are variations in mortality due to causes of death in South Africa's nine provinces.
2. Find out the nature or trend of variation in all-cause mortality in South Africa's nine provinces.

2.2.2 Hypothesis

The multidimensional demographic model describes mortality in terms of events which are simultaneously categorised using three variables namely, province(P) , time(Y) and cause of death(C).

The research hypothesis were:

1. There are variations in mortality due to causes of mortality in South Africa's nine provinces.
2. There are variations in all-cause mortality in South Africa's nine provinces.

Cause of death hypothesis of independence tested was:

$$\begin{aligned} H_0 &: log(M_{ijk}) = \mu + \lambda_i^C + \lambda_j^P + \lambda_k^Y \qquad \text{versus} \\ H_1 &: log(M_{ijk}) = \mu + \lambda_i^C + \lambda_j^P + \lambda_k^Y + \lambda_{ij}^{CP} + \lambda_{ik}^{CY} + \lambda_{jk}^{PY} + \lambda_{ijk}^{CPY} \end{aligned}$$

$\forall\ i = 1, \ldots, 6; \quad j = 1, \ldots, 9; \quad k = 1, \ldots, 11$

where log(M_{ijk}) denoted the logarithm to the base e of the predicted mortality, μ denoted the overall average of the natural logarithm of predicted mortality, λ_i^C denoted the major term of the factor cause of death, λ_j^P showed the major term of the factor province, λ_k^Y denoted the major term of the factor year, λ_{ij}^{CP}, denoted the interaction term for factors cause of death and province, λ_{ik}^{CY} denoted the interaction term for factors cause of death and year, λ_{jk}^{PY} denoted the interaction term for factors province and year, λ_{ijk}^{CPY} denoted the interaction term of the factors cause of death, province and year.

2.3 Research design

2.3.1 Type of data and description of data

Mortality data for South Africa from 2005 to 2015 was used in this research. This data was collected from the statistical releases by Stats SA on mortality and causes of death in South Africa; findings from death notifications. Mortality rates were computed using 2011 Census surveys data for each province also obtained from Stats SA. Direct standardisation was employed to compute the mortality rates. These mortality rates were reported per 1 000 000 people. Log-linear analysis was employed in this study to model mortality data for all the nine provinces of South Africa using statistical packages SPSS and R.

The variables investigated by log-linear analysis are all taken as dependent factors. For the purpose of modeling in this research, the number of deaths (mortality) was taken as the response variable whereas province, year and cause of death are considered as explanatory variables. Since the explanatory factors are categorical and the dependent factor is count, then the dataset qualifies for log-linear analysis. In order to perform log-linear analysis, Poisson distribution was applied with counts as the dependent variable since it is presumed that the dependent variate follows a Poisson distribution.

The multidimensional demographic models describe mortality in terms of events which are simultaneously classified using three variables: province (P), year (Y) and cause of death (C). Since mortality by cause of death forms an integral part of the analysis, the number of causes of death was limited to six categories and if more detailed causes of death are used, then more cells with zero deaths are generated.

The variables used and their categories are summarised in Table 2.1 below.

Table 2.1: Variable categories

Province (P)	**Year (Y)**	**Cause of death (C)**
1. Western Cape (WC)	1. 2005	1. Tuberculosis (TB)
2. Eastern Cape (EC)	2. 2006	2. Diabetes (D)
3. Northern Cape (NC)	3. 2007	3. Various forms of heart diseases (VFHD)
4. Free State (FS)	4. 2008	4. Cerebrovascular (CBV)
5. KwaZulu Natal (KZN)	5. 2009	5. HIV, influenza and pnumonia (HIP)
6. North West (NW)	6. 2010	6. Other natural and non-natural (ONN)
7. Gauteng (GP)	7. 2011	
8. Mpumalanga (MP)	8. 2012	
9. Limpopo (LIM)	9. 2013	
	10. 2014	
	11. 2015	

2.3.2 Sample size

Since log-linear analysis involves MLE as an estimation approach, large samples are appropriate for log-linear analysis (Tabachnick & Fidel, 2014). Samples that are too small may result in problems of low statistical power and on the other hand samples that are too large may also result in problems of unnecessary components being contained in the final model as statistically significant.

In log-linear analysis, sample sizes greater than five times the number of cells in a contingency table will yield accurate results compared to smaller samples (Tabachnick & Fidel, 2014). The sample size for the dataset is appropriate for log-linear analysis since the sample size is well above five times the number of cells in the cross-table which is the minimum sample size to achieve a

high or strong statistical power of .8. For mortality by cause of death dataset, the number of cells is 594 with a sample size of 1 097 283.

2.4 Estimation of the model

Almost all the techniques for examination of fit, selection of the model and interpretation in log-linear model analysis are applicable and theoretically valid if and only if the MLE exists. MLE existence validates the use of large χ^2 approximations to several measures of goodness-of-fit usually applied for model selection and assessment. If the MLE does not exist, the model is unidentifiable since the degrees of freedom tend to be meaningless and the asymptotic standard errors are not defined well. Fienberg and Rinaldo (2012) argued that non-existence of MLE emanates from data lacking full information about parameters of the model and also by sampling zeros.

2.5 Model fit and selection

In log-linear analysis, the model selection process becomes more difficult as the dimension of the table gets bigger and bigger. In model selection procedure, two goals compete: the model should be easy to interpret, smooth rather than over fitting the data. On the other hand, for it to fit the data well, it should be complex. The best model to fit the data is found by following a model selection process which involves a sequential search between hierarchical nested model, beginning with the most possible complex model (saturated model) and taking out interaction effects or terms of less significance one at a time. The process only stops after obtaining a model of best fit. The above described process is commonly known as the backward elimination method. An alternative method to backward elimination method is forward selection. Forward selection starts from the complete independence model and interaction effects are added provided that they significantly improve the fit.

In the case where there is more than one reasonable model, choosing a single best model requires some subjective judgment. Some useful diagnostics tests have been proposed to determine the most appropriate model from the list of potential models. The proposed diagnostics tests include, substantive importance and considerations, closeness between measured and expected odds ratios, correlations between observed and fitted values, analysis of association and information criteria which include Akaike information (AIC), Bayesian information (BIC) and others (Anderson, 2017).

2.5.1 Goodness-of-fit test

Goodness-of-fit test statistic is applied to determine the importance of variates or factors in a model. The goodness-of-fit test is used to measure how well-observed data relates to the fitted model through analysis of deviance of residuals or by comparing fitted values to the measured values.

The corresponding hypothesis tested by the goodness-of-fit test statistic is as follows:

$$\begin{aligned} H_0 &: \text{Model } M_o \text{ fits versus} \\ H_1 &: \text{Model } M_o \text{ does not fit.} \end{aligned}$$

The most commonly used traditional goodness-of-fit test statistics to examine the fit and adequacy of a log-linear model M_0 are Pearson's chi-square (χ^2) and the likelihood ratio statistic (L^2) defined for an $A_1 \times A_2 \times ... \times A_k$ table as:

$$\chi^2 = \sum_{a_1,...,a_k} \frac{(n_{a_1,...,a_k} - \hat{m}_{a_1,...,a_k})^2}{\hat{m}_{a_1,...,a_k}}$$

$$L^2 = 2 \sum_{a_1,...,a_k} n_{a_1,...,a_k} \; log(\frac{n_{a_1,...,a_k}}{\hat{m}_{a_1,...,a_k}}).$$

χ^2 and $L^2 \sim \chi^2_{d-d_0}$ under model M_0 , where $d = \pi_{c=1}^{k} A_c$ -1 is the sum of free cells of the table under consideration, d_0 the number of free coefficients of the assumed model M_0 and $\hat{m}_{a_1}, ..., a_k$ the MLE of the predicted under M_0 frequency for cell $(a_1, ...a_k)$ (Kateri, 2014).

Between these two test statistics, the Pearson's χ^2 statistic is more closer to the chi-square distribution than the likelihood ratio statistic L^2. According to Von-eye Mun (2012), the

likelihood statistic compared to the Pearson's χ^2 statistic has better decomposition characteristics and as a result, it is preferred by researchers to compare hierarchical models and also it is often used for the decomposition of effects in contingency tables.

2.5.2 Information Criteria

Information criteria refers to different types of indices or statistics that are used to compare goodness-of-fit of models to data, number of parameters of the model and size of sample. The models tested by these indices do not have to be nested unlike partial association tests. AIC and BIC are the most frequently employed ones.

2.5.2.1 Akaike Information Criteria (AIC)

Incomplete models in log-linear analysis that somehow attain an acceptable level of goodness-of-fit are normally called parsimonious models. If there are two competing models with the same number of degrees of freedom, then the one having lower order interaction effects is more parsimonious (Von-eye & Mun, 2012). Akaike information criteria explains how much data is lost when a model is fitted. The AIC is expressed as:

$$AIC = -2logL + 2k = 2k - 2logL$$

with k representing number of parameters.

The other formula for calculating AIC is given by:

$$AIC = L^2 - 2d$$

with d being the number of degrees of freedom of the calculated L^2. Below are two variations of the AIC. When the size of the sample increases, AIC is adjusted to AIC_c expressed as:

$$AIC_c = AIC + \frac{2k(k+1)}{n-k-1}$$

and Quasi Akaike Information Criteria given by:

$$QAIC = 2k - \frac{1}{c}\ 2logL_m + \frac{2k(k+1)}{n-k-1}$$

which is responsible for adjusting over dispersion (Von-eye & Mun, 2012).

From a group of possible models to fit the data, the one having the smallest AIC fits the data well (Hawkins, 2014).

2.5.2.2 Bayesian Information Criteria (BIC)

The information criterion which takes the sample size into consideration is the Bayesian information criteria which is expressed as:

$$BIC = -2logB + dlogN = L^2 - dlogN$$

with N being the sum of observations, B posterior odds, L^2 is the likelihood ratio statistic and d is the number of degrees of freedom of the model. When the BIC is negative, M_0 is accepted and is preferable to the complex or saturated model. If there is a group of models being compared, the one having the smallest BIC value is the best model. The BIC process gives a stable model in a way that in large samples, it selects the best model or true model with high likelihood (Anderson, 2017).

3. Results

3.1 Model selection and fitting

Log-linear approach was employed to establish the levels and strengths of association across the response variable number of deaths (mortality) and the independent variables. This technique is preferred over other statistical techniques due to its capability to give useful results and information on multiple associations among numerous variables which are crucial for interpreting the major problem.

Log-linear analysis was performed using R statistical package using the *loglm*() function. Two approaches, Backward Elimination and Forward Selection were used in this study to fit the best model for the data.

3.1.1 Model selection

3.1.1.1 K-way effects and higher order effects

When applying backward elimination technique, there is a table showing k-way terms and higher order terms. This table is divided into two sections: the K-way and Higher order terms and the K-way terms. The first section gives tests of all effects at level k and higher whereas the second section gives tests of just the k-way effects.

The fundamental use of the k-way effects and higher order effects is to inspect model enhancement on inclusion of first order effects, second order effects and so on. A significant p value ($p < .05$) denotes that incorporation or inclusion of the higher order effects enhances the model and the change in the probability ratio value shows how much the enhancement is.

Table 3.1 displays the k-way and higher order effects.

Table 3.1: K-way and higher order effects

K-way and higher order effects	**Likelihood Ratio χ^2**	**Pearson χ^2**	**df**	**P-value**
1	1 429 278.817	2 043 037.679	593	$p < .001$
2	77 328.123	144 286.027	570	$p < .001$
3	26 438.025	27 862.282	400	$p < .001$
K-way effects				
1	1 351 950.695	1 898 751.651	23	$p < .001$
2	50 890.098	116 423.746	170	$p < .001$
3	26 438.025	27 862.282	400	$p < .001$

From Table 3.1, all the 1, 2 and 3 k factors have p-values $< .001$ implying that they are all of high significance. This shows that the major effects, first order and second order interaction terms are supposed to be part of the resultant model. Looking at the table, 1 429 278.817 is the likelihood ratio chi-square of the mean only excluding parameters while 77 328.123 is the likelihood ratio chi-square for the complete independence model. The difference between 1 429 278.817 and 77 328.025 is 1 351 950.695 which weighs the improvement of the model after including main effects. Since the improvement is related with a p-value $< .05$, the hypothesis that main effects do not exist is rejected.

The inclusion of 2-way effects also changes the likelihood ratio chi-square by 50 890.098 with a p-value $<.001$ implying that it is significant. Since the p-value $< .05$, the hypothesis that the first order effects do not exist is rejected. The likelihood ratio chi-square is also improved by 26 438.025 when second order effects are included which is also significant since the p-value $< .001$. The hypothesis that the second order effects are zero is rejected since $p < .05$ meaning that second

order effects exist.

Since it has been established that major effects and both the 2-way and 3-way effects exist, the final model being suggested is saturated model.

3.1.2 Partial associations

Table 3.2 below displays the partial associations.

Table 3.2: Partial associations

Effect	**Partial Chi-Square**	**Number of Iterations**	**df**	**P-value**
Y*C	14 333.741	2	50	$p < .001$
Y*P	11 864.104	2	80	$p < .001$
C*P	21828.530	2	40	$p < .001$
Y	13 752.518	2	10	$p < .001$
C	1 299 252.997	2	5	$p < .001$
P	38 945.180	2	8	$p < .001$

The interaction effects with p-values $< .001$ are significant. From the values displayed in Table 3.2, all the main terms and the 2-way interaction effects are significant implying that the possibility of them being included in the final model is very high.

3.1.3 Backward elimination

Different effects in a sequence are fit in stepwise selection techniques and if an effect's p-value $> .05$, it is removed or eliminated. The starting point of the backward elimination approach needs to be determined and in most cases, it is determined by assessing the quality of the uniform order models which include complete independence model, 2-way interactions (Homogeneous association) model and the 3-way interaction (saturated) model. If the p-value of a log-linear model $>.05$, then the model accounts for the data adequately.

Table 3.3 displays the summary statistics of uniform order models.

Table 3.3: Summary statistics results for uniform order models

Model	**Likelihood Ratio χ^2**	**Pearson χ^2**	**df**	**P-value**
(CPY)	0	0	0	1
(CP)(CY)(PY)	26 438.02	27 862.26	400	$p < .001$
(C)(P)(Y)	77 328.12	144 286.03	570	$p < .001$

Looking at the summary statistics of the uniform order models (Table 3.3) obtained through using the *loglm* () function in MASS library in *R*, (C)(P)(Y), (CP)(CY)(PY) and (CPY), the full model (CPY) is the best starting model.

The 3-way interaction CPY is removed in the next step of backward elimination and the model is assessed for its adequacy. If the significance of the term $> .05$, then it is deleted from the model. The removal of CPY interaction leaves us with the homogeneous association model, that is a model with all the three 2-way interaction effects. Looking at the summary statistics for uniform order models (Table 3.3), this homogeneous model does not adequately fit the data since its likelihood ratio chi-square 26 438.02 corresponds to a p-value of $< .05$.

Table 3.4 displays summary statistics results after removing the interaction effect CPY and one 2-way interaction effect.

Table 3.4: Summary statistics results after removing one 2-way interaction effect and CPY

Model	**Likelihood Ratio χ^2**	**Pearson χ^2**	**df**	**P-value**
(CP)(CY)	38 302.13	45 000.96	480	$p < .001$
(CP)(PY)	40 771.77	51 463.35	450	$p < .001$
(CY)(PY)	48 266.55	56 125.44	440	$p < .001$

Models with two 2-way interaction effects were also investigated to examine their adequacy after deletion or elimination of one 2-way interaction effect and CPY. It can be noted from the summary statistics results (Table 3.4) that all the two 2-way interaction models do not fit the data well since all their likelihood ratios correspond to p-values $< .05$.

Investigating further, the elimination of two 2-way interaction effects and CPY gives us new models and summary statistics results are displayed in Table 3.5. It can be noted that all the models are of poor fit because of the p-values which are all less than .05.

Table 3.5: Summary statistics results after removing two 2-way interaction effects and CPY

Model	**Likelihood Ratio χ^2**	**Pearson χ^2**	**df**	**P-value**
(CP)	67 820.25	101 078.78	540	$p < .001$
(CY)	100 507.7	120 955.9	528	$p < .001$
(PY)	1 363 285	1 870 521	495	$p < .001$

Removal of all the three 2-way interaction effects and CPY effect gives us a complete independence model. From the summary statistics results for uniform order models in Table 3.3, this complete independence model has a likelihood ratio chi-square of 77 328.12 and corresponds to a p-value $< .05$ with 570 df. Since its p-value $< .05$, it means that this model does not adequately fit the data.

Thus the full model is the only one which fits the data adequately. This conclusion is in agreement with the findings from the k-way effects tests that both first order and second order interaction terms have to be incorporated in the final model. They are also supported by the findings from the partial association tests that major effects and first order effects are substantial.

So far, it has been indicated from the partial association and k-way tests that major terms, first order and second order interaction effects should be incorporated in the resultant model. This implies that there is a statistical association between mortality and cause of death, province and year, cause of death and province, cause of death and year and also cause of death, province and year.

3.1.4 Assessing overall fit and diagnostics

Further investigations were done in order to check whether there exist less complex models which adequately fit the data. Goodness-of-fit tests were carried out on three models; full (saturated) model, homogeneous association model and the complete independence model. These models

were selected on the basis of their likelihood ratio statistics. Their likelihood ratio statistics are small compared to the other models. A model having the lowest likelihood ratio is a better model. Also the deviances, AICs and BICs of these three models were compared and the results are displayed in Table 3.6. To determine which of the three models represents the data adequately, the likelihood ratio statistics, AICs, BICs and deviances of these models have to be compared. According to Anderson (2017) and Hawkins (2014), the model with the smallest AIC, BIC, deviance and smallest likelihood ratio fits the data well.

When assessing the fitness of the less saturated models, their residual deviances are compared to the full model. A model fits the data better if its residual deviance is small. The full model (CPY) has a residual deviance of .000 while the homogeneous model (CP)(CY)(PY) changes from the full model by 26 438.025 and the complete independence model (C)(P)(Y) changes the most by 77 328.123. Since the full model has a zero residual deviance, the lowest AIC, BIC and likelihood ratio, it is the most appropriate model.

Table 3.6: Model statistics

Model	**Deviance**	df	**Delta(Dev)**	Delta(df)	**AIC**	BIC	**df**
(CPY)	0.00	0	0.00	0	6 376.308	8 982.115	594
(CP)(CY)(PY)	26 438.02	400	26 438.02	400	32 014.333	32 865.388	194
(C)(P)(Y)	77 328.12	570	50 890.10	170	82 564.43	82 669.72	24

Looking at the model statistics results displayed in Table 3.6, the full model has the least residual deviance of .000, the least AIC of 6 376.308 and the least BIC of 8 982.115. It has also the smallest likelihood ratio of .000 from Table 3.1.

The quality of a model can also be examined by contrasting standardised residuals of the models. Small residuals indicate a good fit and large residuals indicate a poor fit. If the values of the standardised residuals of a model are less than -1.96 or greater than 1.96, then it is an inappropriate model.

Table 3.7: Summary of deviance residuals for uniform order models

Model	**Min**	**1Q**	**Median**	**3Q**	**Max**
(CPY)	0	0	0	0	0
(CP)(CY)(PY)	-41.392	-1.252	.517	2.412	67.577
(C)(P)(Y)	-30.585	-6.391	-.893	3.484	146.092

The complete independence model and the homogeneous model have large standardised residuals and their residual deviances are reflected on the summary statistics in Table 3.7. The minimum and maximum deviance residuals for the homogeneous model are -41.392 and 67.577 respectively. For the complete independence model, the minimum and maximum deviance residuals are -30.585 and 146.092 respectively. Hence these two models do not fit the data well.

Also, if the standardised residual values of a model are all zeros, then the model will fit the data well. From Table 3.7, the full model is having zero standardised residual values showing that it adequately fits the data.

3.1.5 Forward selection

The complete independence model was started with introducing significant effects or terms to the current model. For each model developed, the value of the likelihood ratio chi-square test of fit is indicated and also the significance test of the variance between the new and old model is indicated.

Table 3.8 displays summary statistics results after adding interaction effects CP, CY and PY.

Table 3.8: Summary statistics results after adding interaction effects CP, CY and PY

Model	**LR χ^2**	**Pearson χ^2**	**df**	**Deviance**	**AIC**	**BIC**	**df**	**P-value**
(C)(P)(Y)	77 328.12	144 286.03	570	77 328.12	82 564.43	82 669.72	24	$p < .001$
(CP)(CY)(PY)	26 438.02	27 862.26	400	26 438.02	32 014.33	32 865.39	194	$p < .001$

The three 2-way interaction effects, CP, CY and PY were introduced in the next step of forward selection and the model was assessed for its adequacy. The new model proved to be a better model compared to the independence model. It has a smaller AIC, BIC, likelihood ratio, Pearson Chi-square and deviance values as indicated on the summary statistics Table 3.8 compared to the independence model.

However, this new model does not adequately fit the data since its likelihood ratio chi-square corresponds to a p-value <.05. Since this homogeneous model is of poor fit, the second order interaction effect CPY was added to the homogeneous model giving us a saturated model. The summary statistics results are displayed in Table 3.9 below.

Table 3.9: Summary statistics results after adding CPY to the homogeneous model

Model	**LR χ^2**	**Pearson χ^2**	**df**	**Deviance**	**AIC**	**BIC**	**df**	**P-value**
(CP)(CY)(PY)	26 438.02	27 862.26	400	26 438.02	32 014.33	32 865.39	194	$p < .001$
(CPY)	0	0	0	0.00	6 376.308	8 982.115	594	1

From the summary statistics results displayed in Table 3.9, the full model proved to be a better model compared to the homogeneous model because it has smaller, AIC, BIC, L^2, χ^2 and deviance values compared to the homogeneous model. Since L^2 of the full model corresponds to a p-value $> .05$, then the full model fits the data well.

Since variables cause of death, province and year are independent, the final model of the mortality by cause of death dataset using both backward elimination and forward selection techniques is the saturated model (CPY) given by:

$$log(M_{ijk}) = \mu + \lambda_i^C + \lambda_j^P + \lambda_k^Y + \lambda_{ij}^{CP} + \lambda_{ik}^{CY} + \lambda_{jk}^{PY} + \lambda_{ijk}^{CPY}$$

$\forall\ i = 1, \ldots, 6; \quad j = 1, \ldots, 9; \quad k = 1, \ldots, 11$

3.2 Model parameters

Since the final model has been identified, the next stage was to compute their coefficients. The parameter estimates were computed using the $glm()$ function in the library (vcdExtra) in the R statistical package. This function was preferred over other functions because of its ability to give parameter estimates and extra statistics such as standard errors, z values as well as significance (p) values. All parameter estimates with p-values $<.05$ were significant and were expected to be in the model.

3.2.1 Parameter estimates of the final model

The R package was used to compute the parameter estimates of the model.

3.2.1.1 Parameter estimates and odds ratios for main effects

Table 3.10 displays the coefficients of the main effects computed from the R package. Their standard errors, z values and p values are also displayed. The computed intercept estimate for this dataset is 6.8964 having a standard error of .0016, a z-value of 4 410.813 and a p-value less than .001.

The odds ratios for each parameter estimate are also displayed on Table 3.10. The odds ratios of variable parameters were computed using the formula:

$$OR = exp(variable\ estimate)$$

Table 3.10: Parameter coefficients of the main effects

Main Effect	**Estimate**	**Std. Error**	**Z-value**	**P-value**	**Odds Ratio**
C1	.1060	.0030	34.8062	$p < .001$	1.1118
C2	-.9162	.0045	-202.4245	$p < .001$	.4000
C3	-.4756	.0038	-126.7689	$p < .001$	.6125
C4	-.7036	.0041	-173.3028	$p < .001$	.4948
C5	.0456	.0031	14.7457	$p < .001$	1.0467
P1	-.2425	.0047	-51.9732	$p < .001$	.7847
P2	.0078	.0044	1.7718	$p < .001$	1.0078
P3	.1166	.0042	27.9555	$p < .001$	1.1237
P4	.3420	.0038	88.9767	$p < .001$	1.4078
P5	.1476	.0041	36.3850	$p < .001$	1.1590
P6	.0370	.0046	8.0613	$p < .001$	1.0377
P7	- .1892	.0048	- 394716	$p < .001$	.8276
P8	-.0340	.0045	- 7.6160	$p < .001$	.9666
Y1	.0530	.0053	10.0093	$p < .001$	1.0544
Y2	.1053	.0048	22.0290	$p < .001$	1.1110
Y3	.0896	.0048	18.7287	$p < .001$	1.0937
Y4	.0614	.0049	12.6035	$p < .001$	1.0633
Y5	.0545	.0048	11.2906	$p < .001$	1.0560
Y6	.0366	.0048	7.5841	$p < .001$	1.0373
Y7	-.0009	.0049	-.1800	$p < .001$	.9991
Y8	-.0807	.0050	- 16.1097	$p < .001$	.9225
Y9	-.1008	.0050	-20.1054	$p < .001$	.9041
Y10	-.1060	.0050	-21.1495	$p < .001$	.8994

The first column of the table displays the main effects. C1 stands for the first factor level of variable cause of death which is Tuberculosis, P1 stands for the first factor level of variable province which is Western Cape and Y1 stands for the first factor level of variable year which is 2005. It can be noted from Table 3.10 above that all the main terms are significant at the 5% significance level since their p-values <.05 and hence they are expected in the resultant model.

3.2.2 Parameter estimates for 2-way interaction effects

All the first order association effects were significant at the 5% level and they should be included in the final model. Since they are too many 2-way interaction effects, only a portion of them is displayed below.

Table 3.11: Parameter coefficients of the 2-way interaction effects

Interaction Effect	**Estimate**	**Std. Error**	**Z-value**	**P-value**	**Odds Ratio**
C1P1	-.3968	.0106	37.5459	$p < .001$	.6725
C2P1	.4731	.0116	40.7333	$p < .001$	1.6050
C3P1	.4270	.0099	43.2995	$p < .001$	1.5326
C4P1	.1329	.0120	11.0875	$p < .001$	1.1421
C5P1	-.5843	.0116	-50.5440	$p < .001$	.5575
C1P2	.2123	.0079	26.9817	$p < .001$	1.2365
C2P2	.0475	.0123	3.8762	$p < .001$	1.0486
C3P2	-.2170	.0112	- 19.3562	$p < .001$	.8049
C4P2	.0172	.0113	2.8852	$p < .001$	1.0173
C5P2	-.2086	.0091	- 22.9038	$p < .001$	.8117

The first column displays the two-way interaction effects between variables. C1P1 is the interaction effect between the first factor level of variable cause of death and the first factor level of variable province.

3.2.3 Parameter estimates for 3-way interaction effects

Numerous second order interaction terms were significant at the 5% level. A portion of the 3-way interaction effects are displayed in Table 3.12.

Table 3.12: Parameter coefficients of the 3-way interaction effects

Interaction Effect	**Estimate**	**Std. Error**	**Z-value**	**P-value**	**Odds Ratio**
C1P1Y1	-.4224	.0337	- 12.5489	$p < .001$	.6555
C2P1Y1	.2656	.0400	6.6352	$p < .001$	1.3042
C3P1Y1	.1201	.0306	3.9265	$p < .001$	1.1276
C4P1Y1	.0677	.0380	1.7829	$p < .001$	1.0700
C5P1Y1	.0399	.0359	1.1088	$p < .001$	1.0407
C1P2Y1	-.2071	.0222	- 9.3124	$p < .001$	.8129
C2P2Y1	.2039	.0407	5.0149	$p < .001$	1.2262
C3P2Y1	-.1160	.0352	- 3.2903	$p < .001$	.8905
C4P2Y1	.0095	.0344	.2766	$p < .001$	1.0095
C5P2Y1	.0952	.0263	3.6206	$p < .001$	1.0999

The first column displays the three-way interaction effects among the variables. C1P1Y1 is the interaction effect among the first factor levels of variables cause of death, province and year.

3.2.3.1 Final model

The model of best fit generated is the saturated model given:

$$log(M_{ijk}) = \mu + \lambda_i^C + \lambda_j^P + \lambda_k^Y + \lambda_{ij}^{CP} + \lambda_{ik}^{CY} + \lambda_{jk}^{PY} + \lambda_{ijk}^{CPY}$$

where log(M_{ijk}) denotes the logarithm to the base e of the predicted mortality,

- μ denotes the overall average of the logarithm to the base e of predicted mortality. In this case it is the intercept estimate given by 6.8964.
- λ denotes variable effects.
- i, j, k indicate variable categories.
- λ_i^C denotes the main term of the factor cause of death, λ_j^P shows the major term of the variate province, λ_k^Y denotes the major term of the factor year, $\forall$ i = 1, . . . , 5; j = 1,. . . , 8; k =1, . . . , 10.

- λ_{ij}^{CP}, denotes the interaction term for factors cause of death and province, for all significant (i, j).
- λ_{ik}^{CY}, denotes the interaction term for factors cause of death and year, for all significant (i, k).
- λ_{jk}^{PY}, denotes the interaction term for factors province and year, for all significant (j, k).
- λ_{ijk}^{CPY} denotes the interaction term of the factors cause of death, province and year, for all significant (i, j, k).

Using the parameter estimates obtained from the R package, the generated model is:

$$log(M_{ijk}) = 6.8964 + 0.1060_1^C - 0.9162_2^C - 0.4756_3^C + \ldots, -0.2425_1^P + 0.0078_2^P + \ldots, +0.0530_1^Y + 0.1053_2^Y + \ldots, -0.3968_{1,1}^{CP} + 0.4731_{2,1}^{CP} + \ldots, +0.3922_{1,1}^{CY} - 0.3883_{2,1}^{CY} + \ldots, -0.0797_{1,1}^{PY} + 0.0849_{2,1}^{PY} + \ldots, -0.4223_{1,1,1}^{CPY} + 0.2656_{2,1,1}^{CPY} + \ldots, -0.0122_{5,8,10}^{CPY}.$$

This model is made up of all the main effects, first and second order interaction effects.

4. Discussion

4.1 Model interpretation

For this dataset, log-linear modeling was performed on 3 variables: Cause of death (C), province (P) and year (Y). The results revealed that the full model (CPY) was the best model to fit the data well with the following values of goodness of fit statistics:
Likelihood ratio (L^2) = 0; d.f = 0; p-value = 1.00
Pearson Chi-square (χ^2) = 0; d.f = 0; p-value = 1.00
AIC = 6 376.308; BIC = 8 982.115 and Residual deviance = 0
The model adequacy was confirmed by the above results of the test statistics; AIC, BIC, (L^2), (χ^2) and the Residual deviance.

The odds ratios of variable parameters were computed using the formula:

$$OR = exp(variable\ estimate)$$

Thus the bigger the variable estimate, the bigger the odds ratio and also if the variable estimate is positive then there will be an increase in the expected count and if the variable estimate is negative, then there will be a decrease in the expected count.
Looking at Table 3.10, the last category of each variable was considered the reference category and so its parameter was set to zero.

For the variable cause of death, other natural and non-natural causes (C6), was considered the reference category. The odds ratio for mortality by Tuberculosis (C1) = exp(.1060) = 1.1118 meaning that there was approximately a 11% rise in mortality caused by tuberculosis compared to the number of deaths caused by other natural and non-natural causes. Also the estimated odds ratio for mortality by diabetes was exp(- .9162) = .4000 implying that there was approximately 60% decrease in the number of deaths caused by diabetes compared to other natural and non-natural causes. For the various forms of heart diseases, the odds ratio = exp(-.4756) = .6125 giving an impression that there was approximately 40% decrease in the number of deaths due to various forms of heart diseases. The odds ratio for cerebrovascular = expo(-.7036) = .4948 meaning that there was approximately a decrease of 50% in the number of deaths caused by cerebrovascular compared to other natural and non-natural causes. The odds ratio for HIV, influenza and pneumonia = exp(.0456) = 1.0467 implying that there was approximately 5% increase in the number of deaths caused by HIV, influenza and pneumonia.

From the table of coefficients for main effects, tuberculosis and HIV, influenza and pneumonia have the greatest odds ratio values compared to other causes of death. This means that tuberculosis is one of the top killers in South Africa. These results are supported by the report by Sloane (2018) who reported that tuberculosis remains the main killer in South Africa. Pillay-van Wyk et al. (2016) also reported that HIV and AIDS claimed the majority of the deaths (29.1%) from 1997 to 2012 in South Africa. This is also supported by Dlamini, Goon, Okafor and Mangi (2018) who reported that tuberculosis and HIV and AIDS increased the likelihood of mortality in Zululand District in South Africa.

For the variable province, Limpopo province was considered the reference category. The estimated odds ratio of mortality for Western Cape (P1) was = exp(-.2425) = .7847 implying that mortality in Western Cape decreased by approximately 22% compared to the mortality in Limpopo. For KwaZulu Natal (P5), the overall odds ratio was = exp(.1476) = 1.1590 implying that there was approximately 16% increase in mortality in KwaZulu Natal compared to number of deaths for Limpopo. The number of deaths for Gauteng province was approximately .8276 times more than that of Limpopo and also the number of deaths for Eastern Cape was approximately 1.0078 times more than that of Limpopo. Free State, KwaZulu Natal, Northern Cape, North West and

Eastern Cape recorded the highest number of deaths per 1000 000 since they have the greatest odds ratios compared to other provinces. Western Cape recorded the least mortality rate with an odds ratio of .7847 and Free State recorded the highest mortality rate per 1 000 000 people having an odds ratio of 1.4078. The odds ratio for Western Cape is 1.794 times less than that of Free State. These results are supported by Pillay-van Wyk et al. (2016) who reported that all-cause age standardised death rates were 1.7 times higher in the province with the highest death rate compared to the province with the smallest death rate.

Under variable year, year 2015 was considered the reference category and its parameter was set to zero. Year 2006 had the largest mortality rate followed by year 2007 since they have the highest odds ratios of 1.1110 and 1.0937 respectively. The number of deaths for year 2006 is approximately 1.1110 times greater than that of 2015 and also the number of deaths for year 2007 was approximately 1.0937 times greater than that of 2015. These findings are also supported by Pillay-van Wyk et al. (2016) who pointed out that all-cause mortality levels rose significantly from 1997 and reached the maximum in 2006 and 2007 and then dropped due to improvements in health care facilities.

Since all the 2-way interaction effects were significant, only a few will be highlighted here.

Table 3.11 displays a portion of the 2-way interaction effects for the mortality by cause of death dataset. For the variables cause of death and province, the odds ratio of mortality by diabetes in Western Cape indicates that there were likely to be exp(.4731) = 1.6050 times as frequent. This means that there was approximately an increase of 61% in the number of deaths by diabetes in Western Cape more than expected. The odds ratios of mortality by various forms of heart diseases in Western Cape and Eastern Cape are exp(.4270) = 1.5326 and exp(- .2170) = .8049 respectively. Thus there was approximately a 53% rise in mortality caused by various forms of heart diseases in Western Cape than expected and approximately a 20% decrease in the mortality caused by various forms of heart diseases in Eastern Cape. Using these odds ratio values, it can be deduced that mortality varies among provinces due to causes of death. This is supported by Rademeyer (2017) who reported that, in South Africa, socio-demographic factors as well as the province of residence contribute significantly to mortality. This is also supported by

Made, Wilson, Jina, Tlotleng, Jack, Ntlebi and Kootbodien (2017) who reported that 8% of the deaths in South Africa in 2014 were attributed to cancer with Western Cape having the highest age standardised cancer mortality followed by Northern Cape, with the lowest rate in Limpopo.

For the variables cause of death and year, the interaction between diabetes and year 2014 has a parameter estimate of .2429, the interaction between tuberculosis and year 2008 has a parameter estimate of .1842 and the interaction between various forms of heart diseases and year 2007 has a parameter estimate of - .0080. The odds ratios for these interaction effects are exp(0.2429) =1.2749, exp(.1842) = 1.2023 and exp(- .0080) = .9920 respectively. Thus there was approximately 27% rise in mortality caused by diabetes in 2014 than expected, 20% rise in mortality caused by tuberculosis in year 2008 than expected and also approximately 1% decrease in mortality caused by various forms of heart diseases in 2008 than expected.

For the variables province and year, the interaction between Northern Cape province and year 2014 has a parameter estimate of .1825, the interaction between North West province and year 2006 has a parameter estimate of .1370 and the interaction between KwaZulu Natal province and year 2005 has a parameter estimate of .0549. The odds ratios corresponding to these interaction effects are 1.2002, 1.1468 and 1.0564 respectively. This implies that there was approximately 20% rise in mortality in Northern Cape in 2014 than expected, 15% rise in mortality in Northwest in 2006 than expected and 6% rise in mortality in KwaZulu Natal in 2005 than expected.

All the second order interaction terms were significant at the 5% level. Only the first three interaction parameter estimates are highlighted here. A portion of the 3-way interaction effects are displayed in Table 3.12.

The 3-way interaction interaction between tuberculosis, Western Cape and year 2005 has a parameter estimate of - .4224. The odds ratio for mortality of this interaction indicates that there were likely to be exp(- .4224) = .6555 times as frequent. Thus there was approximately 34% decrease in mortality caused by tuberculosis in Western Cape province in 2005 than expected.

The 3-way interaction effect between diabetes, Western Cape province and year 2005 has a parameter estimate of .2656. The odds ratio of mortality for this interaction effect indicates that there are likely to be exp(.2656) = 1.3042 times as frequent. Thus, there was approximately 30% increase in the number of deaths for diabetes in Western Cape in 2005 than expected.

The interaction between various forms of heart diseases, Western Cape province and year 2005 has a parameter estimate of .1201 and odds ratio of exp(.1201) = 1.1276. Thus the odds ratio for mortality by various forms of heart diseases in Western Cape province in 2005 indicates that there were likely to be 1.1276 times as frequent. This means that there was likely to be an increase of 13% in mortality in Western Cape province due to various forms of heart diseases in 2005 than expected.

The major findings of the analysis reveal that mortality varies significantly by causes of death. It was also revealed that mortality varies widely across the country and among provinces.

4.2 Research limitations

- Causes of death were limited to six categories out of a numerous number of causes of death. This might result in loosing information on other variables.
- Techniques employed for parameter estimation may contribute to some shortfalls which might affect the accuracy of the results. For example, the $glm()$ function used to estimate the parameters uses a default coding scheme which takes the first or last category of a variable as a reference level. As a result, the number of categories of each variable are reduced by one and parameter estimates of these reference levels are not reflected.

4.3 Areas for further study

A lot of statistical techniques was employed in studies related to mortality modeling. Very few have used graphical and decomposable log-linear models to model mortality data.

This research was done in a developing country. It could be replicated in another developing country using latest data to see if it will yield similar results.

Most of the researches done on mortality modeling are based on single populations. A limited number of studies have looked at multi-population mortality models. More studies based on multi-population could be carried out.

5. Conclusion

5.1 Research summary

In this research, log-linear analysis technique was found to be the most appropriate approach to perform analysis of mortality data since the data is discrete and categorical. This technique is a broadly applied statistical tool for qualitative data in multi-way tables. It has the ability to tease out associations among the variables in multi-way tables.

The main focus of this research was to find out whether or not there are variations in mortality due to causes of death and also to find out the nature or trend of variation in all-cause mortality in South Africa's nine provinces. Statistical packages namely SPSS and R were used to perform log-linear analysis applying backward elimination and forward selection approaches to select a model of best fit.

5.2 Research findings

The study aimed at identifying underlying trends in mortality due to causes of death in South Africa.

Log-linear modeling was performed on 3 variables: Cause of death (C), province (P) and year (Y).

The results have shown that the full model (CPY) was the model of best fit. The model adequacy was confirmed by the following test statistics; AIC, BIC, (L^2), (χ^2) and the Residual deviance.

The results showed that mortality is influenced by causes of death. Tuberculosis claimed the highest number of lives followed by HIV, influenza and pneumonia from the period 2005 to 2015 in South Africa. Mortality by various forms of heart diseases was most prevalent in Western Cape and also in Gauteng. Thus, it can be deduced that mortality varies among provinces due to causes of death.

Since there were significant or notable variations between provinces in mortality rates for specific causes of death, the hypothesis that there are variations in mortality due to causes of death in South Africa was not rejected.

It was also discovered that the percentage distribution of mortality due to all causes was not the same in all the provinces from 2005 to 2015.

References

1. Adarabioyo, M.I. 2014. Application of log-linear model to determinants of child mortality in Nigeria. *International journal of science research and innovative technology*, 1(1): 34-44.

2. Alicandro, G., Sebastiani, G., Bertuccio, P., Zengarini, N., Costa, G., La Vecciah, C. and Frova, L. 2018. The main causes of death contributing to absolute and relative socio-economic inequality in Italy. *Journal of Public Health*, 164(2018): 39-48. [Online]. Available at: https://www.sciencedirect.com

3. Anderson, C.J. 2017. Log-linear models for contingencytables [Edpsy/Psych/Soc 589 Lecture Note]. University of Illinois; Department of educational psychology: Spring.

4. Chavhan, R.N. and Shinde, R.L. 2016. Modeling and forecasting mortality using the Lee-Carter model for Indian population based on decade-wise data. *Sri Lankan journal of applied statistics*, 17(1): 51-68.

5. Chikobvu, D. and Shoko, C. 2018. A Markov model to estimate mortality due to HIV/AIDS using CD4 cell counts based states and viral load: A principal component analysis approach. *Biomedical Research*, 29(15):3090-3098. [Online]. Available at: https://www.biomedres.info

6. Dlamini, K.Z., Goon, D.T., Okafor, U.B. and Mangi, N.G. 2018. Factors contributing to mortality among new tuberculosis patients in the Zululand health district, South Africa: A retrospective study. *African Journal for Physical Activities and Health Sciences*, 24(4): 614-626.

7. Fienberg, S. and Rinaldo, A. 2012. Maximum Likelihood Estimation in Log-linear Models. *The Annals of Statistics*, 40(2): 996-1023.

8. Hair, J.F. Jr, Black, W.C., Babin, B.J. and Anderson, R.E. 2019. *Multivariate data analysis*. 8^{th} ed. Hampshire: Cengage Learning.

9. Hawkins, D. 2014. *Biomeasurement : A student's guide to biostatistics*. 3^{rd} ed. Oxford: Oxford University Press.

10. Johnson, L.F., May, M.T., Dorrington, R.E., Cornell, M., Cornell, M., Boulle, A., Egger, M. and Davies, M.A. 2017. *Estimating the impact of antiretroviral treatment on adult mortality trends in South Africa : A mathematical modelling study*, 12 December 2017. [Online]. Available at: https://journals.plos.org/plosmedicine/article?id=10.1371/journal.pmed.1002468

11. Karimi, M., Rey, G. and Latouche, A. 2017. A joint modelling of socio-professional trajectories and cause-specific mortality. *Computational Statistics and Data Analysis*, 119(2018): 39-54.

12. Kateri, M. 2014. *Contingency table analysis : methods and implementation using R, statistics for industry and technology*. New York: Springer Science + Business Media.

13. Lacobucci, D. and Ann, L. McGill. 1990. Analysis of attribution data: Theory testing and effects estimation. *Journal of personality and social psychology*, 59(3): 426-441.

14. Li, H., O'Hare, C. and Zhang, X. 2015. A semiparametric panel approach to mortality modeling. *Insurance : Mathematics and Economics*, 61(2015): 264-270.

15. Li, H., Li, H., Lu, Y. and Panagiotelis, A. 2019. A forecast reconciliation approach to cause-of-death mortality modeling. *Insurance : Mathematics and Statistics*, 86(2019): 122-133.

16. Made, F., Wilson, K., Jina, R., Tlotleng, N.,Jack, S., Ntlebi, V. and Kootbodien, T. 2017. Distribution of cancer motarlity rates by province in South Africa. *Cancer Epidemiology*, 51(2017): 56-61.

17. Manda, S.O.M. and Abdelatif, N. 2017. Smoothed Temporal Atlases of Age-Gender All-Cause Mortality in South Africa. *International journal of environmental research and public health*, 14(1072): 1-18. [Online]. DOl: 10.3390/ijerph14091072

18. Manzo, G. 2014. *Data, generative models and mechanisms : more on the principles of analytical sociology*. Chichester: John Wiley & Sons.

19. Musenge, E., Chirwa, T.B., Kahn, K. and Vounatsou, P. 2012. Bayesian analysis of zero inflated spatiotemporal HIV/TB child mortality data through the INLA and SPDE approaches: Applied to data observed between 1992 and 2010 in rural North East South Africa. *International Journal of Applied Earth Observation and Geo − information*, 22(2013): 86-98.

20. Nyman, H., Pensar, J., Koski, T. and Corander, J. 2015. Context-specific independence in graphical log-linear models. *Article in computational statistics*, 31(4): 1-19. [Online]. DOl: 10.1007/s00180-015-0606-6.

21. Oldenburg, C.E., Seage, G.R., Tanser, F., De Gruttola, V., Mayer, K.H., Mimiaga, M.J., Bor, J. and Barnighausen, T. 2018. Antiretroviral therapy and mortality in rural South Africa: A comparison of casual modeling approaches. *American Journal of epidemiology*, 187(18): 1772-1779. [Online]. DOl: 10.1093/aje/kwy065.

22. Olmus, H. and Erbas, S. 2012. Analysis of Traffic accidents caused by drivers by using log-linear models. *Transportation research board*, 24(6): 495-504.

23. Phetlhu, D.R., Bimerew, M., Marie-Modeste, R.R., Naidoo, M. and Igumbor, J. Nurses' Knowledge of Tuberculosis, HIV, and Integrated HIV/TB Care Policies in Rural Western Cape, South Africa. *Journal of the Association of in Aids Care*, 29(6): 876-886.

24. Pillay-van Wyk, V., Msemburi, W., Laubscher, R., Dorrington, R.E., Groenewald, P., Glass, T., Nojilana, B., Joubert, J.D., Matzopoulos, R., Prinsloo, M., Nannan, N., Gwebushe, N., Vos, T., Somdyala, N., Sithole, N., Neethling, I., Nicol, E., Rossouw, A. and Bradshaw, D. 2016. Mortality trends and differentials in South Africa from 1997 to 2012: second National Burden of Disease Study. *Burden of disease research*, 4: 642-653. [Online]. Available at: https://www.thelancet.com/lancetgh

25. Rademeyer, S. 2017. Provincial differentials in under- five mortality in South Africa. Master in Population Studies dissertation. University of Kwazulu Natal.

26. Ritchie, H. and Roser, M. 2018. *Causes of death*, February 2018. [Online]. Available at: https://www.ourworldindata.org/causes of death

27. Scovronick, N., Sera, F., Acquaotta, F., Garzena, D., Fratianni, S., Wright, C.Y. and Gasparrini, A. 2017. The association between ambient temperature and mortality in South Africa: A time-series analysis. *Journal on Environmental Research*, 161(2018): 229-235.

28. Sloane, H. 2018. *Top* 10 *Leading Causes of Death in South Africa*, 03 April 2018. [Online]. Available at: https://www.2oceansvibe.com/2018/04/03/top-10-leading-causes-of-death-in-south-africa/ https://www.cambridge-news.co.uk/news/leading-death-dementia-diseases-133300775 [Accessed 01 November 2018].

29. Statistics South Africa [Stats SA]. 2005-2017. *Mortality and causes of death in South Africa : Findings from death notification forms.* Pretoria: Statistics South Africa.

30. Tabachnick, B.G. and Fidell, L.S. 2014. *Using multivariate statistics.* 6^{th} ed. Boston, MA: Pearson.

31. Udjo, E.O. and Lalthapersad-Pillay, P. 2013. Estimating maternal mortality and causes in South Africa: National and provincial levels. *Journal on Midwifery*, 30(2014): 512-518.

32. United Nations. 2014. *Principles and recommendations for vital statistics system, third revision*: Department of economic and social affairs. New York: United Nations publications.

33. Von-Eye, A. and Mun, E. 2012. *Log−linear modeling : Concepts, Interpretation and Application.* New Jersey: John Wiley & Sons.

34. Yifang Han. 2013. The analysis of child mortality in a developing country. Master thesis. Stockholm University.