

**BAYESIAN HIERARCHICAL MODELS WITH APPLICATIONS TO
CERVICAL, OESOPHAGEAL AND LUNG CANCERS IN
KENYA'S COUNTIES**

**BY
JOSEPH KURIA WAITARA**


**A THESIS SUBMITTED TO THE SCHOOL OF SCIENCES AND
AEROSPACE STUDIES IN PARTIAL FULFILLMENT OF THE
REQUIREMENTS FOR THE AWARD OF DOCTOR OF
PHILOSOPHY IN BIO-STATISTICS**

MOI UNIVERSITY

2022

DECLARATION


I do hereby declare that this Thesis is my original work and has never been printed or published anywhere or in any institution of higher learning for purposes of academic accreditation.

Signature  Date 19-11-2022

Joseph Kuria Waitara

DPS/PHD/002/2017


This Thesis has been submitted for review with our approval as University supervisors.

Signature  Date 19-11-2022

Dr. G. Kerich

Mathematics, Physics and Computing Department

Moi University

Signature  Date 19-11-2022

Prof. J. Kihoro

Computing and Mathematics Department

The Co-operative University of Kenya

DEDICATION

I dedicate this Thesis to my dear family for support and encouragement they have offered throughout my education.

ACKNOWLEDGMENTS

I would like to express sincere gratitude to my supervisor Dr. G. Kerich for great guidance towards the preparation of this Thesis. Special thanks to my supervisor Prof. J. Kihoro for outstanding support you offered throughout Thesis development. I express my heartfelt appreciation Dr. Matthew Kosgei, Moi University, Ms. Ann Korir, Evans Kiptanui and Nathan Okerosi of National Cancer Registry -Kenya for providing data used in this Thesis. I thank Dr. Ngugi Mwenda who constantly checked on and encouraged me to stay focused until completion. I express sincere appreciation to my family members led by dad Peter, mum Teresia, my lovely wife Sarah, brother Simon, brother John and sisters Margaret and Beth for their endless motivation and support while undertaking my Doctoral studies.

ABSTRACT

Cancer is an event associated with space and time. Counties relative risks estimates can be obtained using Bayesian hierarchical models. The general objective of the research was to obtain county based estimates through Bayesian hierarchical modeling of cervical, oesophageal and lung cancers in Kenya's counties from 2015 to 2016, period which complete data was available. Specific objectives were: to model over-dispersion and conduct spatial correlations tests in order to model three cancer cases distribution in Kenya' counties; to model cervical cancer cases using Poisson-Gamma and spatial-temporal models; to model the effects of co-variates on spatial-temporal distribution of oesophageal and lung cancer cases in Kenya's counties. The data was obtained from National Cancer Registry (NCR) which carried a 2-year retrospective study in ten counties. Cervical cancer cases were 1064, oesophageal cancer cases 1599 while lung cancer cases were 256. A simple Poisson log-linear model dispersion parameter for cervical was 31.202, oesophageal was 49.241 and lung cancer cases was 6.134 which were greater than 1 indicating over dispersion. Spatial correlation tests conducted for the three cancers revealed that there was no spatial auto correlations of the residuals since for cervical cancer $p\text{-value}=0.2104 > 0.05$, oesophageal $p\text{-value}=0.4155 > 0.05$ while for lung cancer $p\text{-value}=0.4120 > 0.05$. The model revealed that the highest cervical cancer relative risk was in Embu=7.92 and lowest in Bomet which was 1.53. The smoking and alcohol use interaction oesophageal cancer model revealed that Bomet=11.16 had the highest risk while Kiambu had the lowest relative risk 0.6. Smoking and alcohol use were significant risk factors of oesophageal cancer. The multiplicative effect of smoking was 1.012, thus 1.2 % higher to smokers compared to non-smokers. Alcohol use was 1.0346 thus 3.5 % higher to alcohol users. The interaction model revealed that oesophageal cancer was 16.88 % higher to alcohol users while it was 4.60 % higher to smokers. The interaction model for lung cancer revealed that in Nairobi=5.97 had highest risk while lowest in Kakamega=0.1. In the lung cancer model the multiplicative effect of smoking was 1.4021, indicating 40.21 % higher to smokers as compared to non-smokers, 1.3689 for alcohol use variable that is 36.89 % higher to alcohol users. In interaction model the effect was 7.86 times higher for smokers. In conclusion, simple Poisson regression models were not appropriate to model the three cancers due to over dispersion nature of the data sets. The spatial correlation tests revealed that there was no spatial auto correlation for the three types of cancer. Application of Bayesian hierarchical models enabled generation of relative risks and identification of the risk patterns of various counties, a major milestone since previous studies focused on specific regions. We recommend that, since all counties had cervical cancer relative risk greater than 1, step up screening and avail vaccines to the appropriate groups. To mitigate oesophageal cancer, counties should create awareness on effects of smoking and alcohol use. In case of lung cancer, counties with relative risks greater than 1 should disseminate information elaborating the effects of smoking and alcohol use.

TABLE OF CONTENTS

1	INTRODUCTION	1
1.1	Background of the study	1
1.2	Statement of the problem	3
1.3	Justification of the study	4
1.4	Objectives	4
1.4.1	General objective	4
1.4.2	Specific objectives	4
1.5	Research Questions and Hypotheses	5
1.6	Significance of the study	6
1.7	Delimitation of the study	6
1.8	Limitation	6
2	LITERATURE REVIEW	7
2.1	Introduction	7
2.2	To model over-dispersion and conduct spatial correlations tests for cervical, oesophageal and lung cancer cases distribution in Kenya's counties.	7
2.3	To model cervical cancer cases using Poisson-Gamma and spatial-temporal models.	10
2.4	To model the effects of covariates on spatial-temporal distribution of oesophageal and lung cancer cases in Kenya's counties.	12
3	METHODOLOGY	17
3.1	Introduction	17
3.2	Methodologies for the Objectives	17
3.2.1	To model over-dispersion and conduct spatial correlations tests in order to model cervical, oesophageal and lung cancer cases distribution in Kenya's counties.	17

3.2.2	To model cervical cancer cases using Poisson-Gamma and Spatial-temporal models.	19
3.2.3	To model the effect of covariates on spatial distribution of oesophageal and lung cancer cases in Kenya's counties . . .	32
3.3	Model selection criteria	34
3.3.1	Deviance Information Criteria	34
3.4	Ethical Approval	34
3.5	Sample and Sampling Technique	35
3.6	Data Analysis	35
4	RESULTS AND DISCUSSION	36
4.1	Introduction	36
4.2	To model over-dispersion and conduct spatial correlations tests in order to model cervical, oesophageal and lung cancer cases distribution in Kenya's counties.	36
4.2.1	Assessing the presence of over-dispersion and spatial correlation for cervical cancer cases	36
4.2.2	Assessing the over dispersion and spatial correlation of oesophageal cancer cases	40
4.2.3	Assessing the over dispersion and spatial correlation of lung cancer data	41
4.3	To model cervical cancer cases using Poisson-Gamma and Spatial-temporal models.	42
4.3.1	Standard Incidence Ratio (SIR) map for cervical cancer cases	42
4.3.2	Poisson-gamma model	43
4.3.3	Spatial-temporal models for cervical cancer cases	44
4.4	To model the effects of covariates on spatial-temporal distribution of oesophageal and lung cancer cases in Kenya's counties.	48
4.4.1	Descriptive statistics for oesophageal cancer	48
4.4.2	Standardized Incidence Ratio (SIR) of oesophageal cancer .	48

4.4.3	Spatial-temporal models for oesophageal cancer	50
4.4.4	Descriptive statistics for lung cancer	58
4.4.5	Spatio-temporal models for lung cancer	58
5	SUMMARY, CONCLUSION AND RECOMMENDATIONS	68
5.1	Introduction	68
5.2	Summary	68
5.3	Recommendations	70
5.3.1	Recommendations for national and counties governments . .	70
5.3.2	Recommendation for National Cancer Registry and other cancer registries	70
5.3.3	Areas of further research	71
	References	72
	A APPENDICES	79

LIST OF FIGURES

1.1	Kenya Counties	3
4.1	Histogram of cervical cancer cases.	37
4.2	Standardized Incidence Ratio (SIR).	43
4.3	Distribution of the county specific relative risks of cervical cancer in the disease mapping model.	47
4.4	Map of the uncertainty for the spatial effect $\zeta_i : p(\zeta_i > 1 y)$	47
4.5	Standardized Incidence Rates (SIR) for oesophageal cancer.	49
4.6	Spatial-temporal distribution of the relative risks for oesophageal cancer with smoking as the covariate.	54
4.7	Map of the uncertainty for the spatial temporal effects accounting for smoking effect (oesophageal cancer) $\mu_i : p(\mu_i > 1 y)$	54
4.8	Spatial-temporal distribution of the relative risks for oesophageal cancer with alcohol use and smoking as the covariates.	57
4.9	Spatial-temporal distribution of the relative risks for lung cancer with smoking as the covariate.	61
4.10	Map of the probability for the spatial temporal effects accounting for smoking effect (lung cancer) $\mu_i : p(\mu_i > 1 y)$	61
4.11	Spatial-temporal distribution of the relative risks for lung cancer with alcohol use as the covariate.	64
4.12	Map of the probability values accounting for alcohol use (lung cancer) $\mu_i : p(\mu_i > 1 y)$	64
4.13	Spatial-temporal distribution of the relative risks for lung cancer with alcohol use-smoking as the covariate.	67

LIST OF TABLES

1	Poisson log normal model for cervical cancer cases	38
2	Over dispersion test for cervical cancer cases	38
3	Moran I Statistics for cervical cancer cases	39
4	Poisson log normal model for oesophageal cancer cases	40
5	Over dispersion test for oesophageal cancer cases	40
6	Moran I Statistics for oesophageal cancer cases	41
7	Poisson log normal model for lung cancer cases	41
8	Over dispersion test for lung cancer cases	42
9	Moran I Statistics for oesophageal cancer cases	42
10	Poisson-Gamma model: Summary of the fixed effects estimates . . .	44
11	Relative risks for cervical cancer Poisson-Gamma model . . .	44
12	Deviance Information Criterion (DIC) for the three Spatial-temporal models	45
13	Relative risks for cervical cancer spatial temporal model . . .	46
14	Distribution of oesophageal cancer by gender in 2015	48
15	Distribution of oesophageal cancer by gender in 2016	48
16	Oesophageal cancer Standardized Incidence Ratios (SIR)	49
17	Results for various oesophageal models fitted with smoking as the covariate	51
18	The relative risks for counties with notified oesophageal cancer cases where smoking was the covariate	51
19	Results for various models fitted with alcohol use as the covariate . .	52
20	The relative risks for counties with notified oesophageal cancer cases with alcohol use as the covariate	53
21	Results for various oesophageal models fitted with Alcohol, Smok- ing, Year and an interaction as the covariate	55
22	The relative risks for Model 3 where alcohol use and smoking were the covariates	56

23	Distribution of lung cancer by gender in 2015	58
24	Distribution of lung cancer by gender in 2016	58
25	Results for various models fitted with smoking as the covariate . . .	59
26	The relative risks for counties with notified lung cancer cases with smoking as the covariate	59
27	Results for various models fitted with alcohol use as the covariate .	62
28	The relative risks for counties with notified lung cancer cases where alcohol use is the covariate	63
29	Results for various lung cancer models fitted with alcohol use, smok- ing, year and an interaction as the covariate	65
30	The relative risks for counties with notified lung cancer cases where smoking and alcohol use were the covariates	66

NOMENCLATURE

BYM	:Besag,-York-Mollié
CAR	: Conditional Auto-regressive Priors
DIC	: Deviance Information Criterion
GLMM	: Generalized Linear Mixed Model
INLA	:Integrated Nested Laplace Approximation
MCMC	: Markov chain Monte Carlo
SAE	: Small Area Estimation
SIR	:Standardized Incidence Ratio
SPDE	:Stochastic Partial Differential Equation

CHAPTER ONE

1 INTRODUCTION

This chapter will focus on introduction, background of the study, statement of the problem, justification of the study, hypotheses and the research objectives.

1.1 Background of the study

Cancer is a generic term for a large group of diseases characterized by the growth and spread of abnormal cells beyond their usual boundaries that can then invade adjoining parts of the body and/or spread to other organs. Cancer arises when normal cells transform into tumor cells in various stages from pre-cancerous lesion to a malignant tumour. According to the International Agency for Research on Cancer (IARC), the global cancer burden is estimated to have risen to 18.1 million new cases and 9.6 million deaths in 2018. One in 5 men and one in 6 women worldwide develop cancer during their lifetime, and one in 8 men and one in 11 women die from the disease. In Africa, Cancer is an emerging health problem where in 2012 new cancer cases were about 847,00 and 519,00 deaths, three quarters of those deaths occurred in sub-Saharan region (Parkin et al., 2014).

In Kenya cancer cases and mortality rate has risen drastically to worrying levels. In 2018, cancer was ranked as the third leading cause of deaths in Kenya after infectious and cardiovascular diseases. The annual incidence of cancer in Kenya was estimated to be 47,887 new cancer cases, with an annual mortality of 32,987 in 2018. Among men, prostate, oesophageal and colorectal were the leading cancers, while among women, breast, cervical and oesophageal cancers were most common. The leading cause of cancer death in Kenya was oesophageal cancer contributing 13.2 % (4,351 deaths) of cancer mortality. According to data from World Health Organization in 2020, in Kenya breast cancer was the most prevalent followed by cervical cancer among women.

Studies to analyze its dynamics may aid in implementation of strategies and interventions to modify risk behaviors among the people. Mapping cancer rates is important for visualizing spatial or spatial-temporal patterns that may help identify differences in disease burden for different geographic locations while locating areas with high rates of cancer. Identification of high disease burden areas will help in prioritization of cancer control efforts and strategies. According to the International Agency for Research on Cancer Bray et al. (2018), 11 infectious agents have been classified and established as carcinogenic agents in humans namely: *Helicobacter pylori*, Hepatitis B Virus (HBV), Hepatitis C virus (HCV), Human Immunodeficiency Virus type 1 (HIV-1), Human Papillomavirus (HPV), Epstein-Barr virus (EBV), Human Herpes Virus type 8 (HHV-8; also known as Kaposi's sarcoma herpes virus), Human T-cell Lymphotropic Virus type 1 (HTLV-1), *Opisthorchis viverrini*, *Clonorchis sinensis*, and *Schistosoma haematobium*. Other virus-associated cancers include: Kaposi's sarcoma (caused by human herpes virus 8), adult T-cell leukemia (caused by human T-cell leukemia virus type 1), and lymphomas caused by Epstein Barr virus (Epstein et al., 1964).

Kenya is divided into 47 administrative units (See Figure 1.1) referred to as counties and covers an area of 582,650 km square with a population of 47.5 Million people in 2019. The study sought to create a spatial-temporal model to analyze the spatial dynamics of cancer cases in Kenya from data obtained in two years (2015 and 2016) by Nairobi Cancer Registry (NCR).



Figure 1.1: Kenya Counties

1.2 Statement of the problem

Events that occur anywhere are associated with location and time thus spatial and temporal components of these events can be combined to demonstrate aspects related to when and where these events occurred. Cancer is also an event associated with space and time therefore analysis can be done on cancer data to determine its spread, patterns and trends to come up with ways to halt its spread (Mwangi, 2014). Kenya is lagging behind due to lack of nationwide data for all cancer cases since it has only two cancer registries namely; Nairobi Cancer Registry (NCR) and Eldoret Cancer Registry (ECR). This has grossly hindered the cancer prevalence studies (Mwangi, 2014). Non-availability of substantive county based data hinders awareness programs, establishment and accessibility of cancer screening

and treatment facilities. Incidence rates available does not reflect county based estimates they are inadequate and non-informative on distribution of cancer cases in counties in Kenya. Availability of county-based estimates and spatial-temporal distribution of cancer cases will aide development of targeted county strategies, promote awareness, and implementation of universal coverage of key cancer control interventions which will be vital in halting and reducing the rising burden of cancer in Kenya.

1.3 Justification of the study

According to Mathers et al. (2005) national estimates of cancer incidence and mortality are predominantly based on data from population-based cancer registries (PBCR), most of which cover relatively limited sub-national populations. Cancer prevalence studies require detailed data from all over the country about the cancer types, age, gender, location and status of health facilities dedicated to cancer cases. The study sought to develop appropriate spatial-temporal models to provide estimates of the distribution of cancer cases in counties neighbouring the counties where the cancer data is available.

1.4 Objectives

1.4.1 General objective

The general objective of the research was to obtain county based estimates through Bayesian hierarchical modeling of cervical, oesophageal and lung cancers in Kenya's counties from 2015 to 2016.

1.4.2 Specific objectives

1. To conduct over-dispersion test and model spatial correlations for cervical, oesophageal and lung cancer cases distribution in Kenya's counties.

2. To model cervical cancer cases using Poisson-Gamma and Spatial-temporal models.
3. To model the effects of covariates on spatial-temporal distribution of oesophageal and lung cancer cases in Kenya's counties.

1.5 Research Questions and Hypotheses

1. Hypotheses for the first objective

H_0 : Cervical, oesophageal and lung cancer cases are not over dispersed.

H_1 : Cervical, oesophageal and lung cancer cases are over dispersed.

H_0 : Cervical, oesophageal and lung cancer cases are randomly dispersed.

H_1 : Cervical, oesophageal and lung cancer cases are not randomly dispersed.

2. What are the relative risks of cervical cancer cases in counties in Kenya?

- 3(a) Hypotheses for oesophageal cancer

H_0 : Smoking is not a risk factor for oesophageal cancer.

H_1 : Smoking is a risk factor for oesophageal cancer.

H_0 : Alcohol use is not a risk factor for oesophageal cancer.

H_1 : Alcohol use is a risk factor for oesophageal cancer.

H_0 : Smoking is not a significant risk factor for oesophageal cancer when controlling for alcohol use.

H_1 : Smoking is a significant risk factor for oesophageal cancer when controlling for alcohol use.

- 3 (b) Hypotheses for lung cancer.

H_0 : Smoking is not a risk factor for lung cancer.

H_1 : Smoking is a risk factor for lung cancer.

H_0 : Alcohol use is not a risk factor for lung cancer.

H_1 : Alcohol use is a risk factor for lung cancer.

H_0 : Smoking is not a significant risk factor for lung cancer when controlling for alcohol use.

H_1 : Smoking is a significant predictor for lung cancer when controlling for alcohol

use.

1.6 Significance of the study

A better characterization of the county differences in cancer distribution would guide the awareness campaign, detection programs and enable effective treatment of detected cases reducing cancer mortality rates and suffering of the ailing patients. The study sought to add to the existing body of knowledge which would inform the National Government and County Governments in policy formulation to address the cancer burden in Kenya.

1.7 Delimitation of the study

This study was carried out using cancer cases data available in some counties in Kenya. The data was obtained from National Cancer Registry which in 2015 and 2016 conducted a county based surveillance study of all reported cancer cases in hospitals in various counties in Kenya.

1.8 Limitation

In this study, main limitation was obtaining adequate data since data available was for ten counties out of forty seven.

CHAPTER TWO

2 LITERATURE REVIEW

2.1 Introduction

This chapter provides the literature in relation to over-dispersion model, spatial correlation test and Bayesian hierarchical models. Further, it provides literature of the previous studies on cervical, oesophageal and lung cancers.

2.2 To model over-dispersion and conduct spatial correlations tests for cervical, oesophageal and lung cancer cases distribution in Kenya's counties.

The number of occurrence of any event within a specified time can be described as counting data. In the case of the dependent variable is a count and researcher is interested in how this count changes as the explanatory variable increases count data regression model is used (Esin, 2018). Counts are positive and for rare events the Poisson distribution instead of normal distribution is appropriate. Simple Poisson log-linear model has been applied widely to model count data. It is suitable for modeling equi-dispersed (i.e an equal mean and variance) distribution. However, the model is not applicable to data set which contains substantial over dispersion. According to Esin (2018) in many instance real data do not adhere to this assumption (over- or under-dispersed data) and the inappropriate imposition of Poisson regression model may underestimate the standard errors and overstate the significance of regression coefficients. Therefore, assessment of presence of over or under dispersion in the data set should be conducted by first computing the residuals from a simple Poisson log-linear model before proceeding to model the diseases.

Surveys are designed for obtaining reliable estimates in the whole population or

in some sub-populations called planned domains. However, it is quite common in practice to use survey data for estimating indicators of non-planned domains (small areas) with small samples sizes. There is a growing demand for estimates for smaller areas or domains. Such estimates are now routinely calculated using the so-called indirect or model-based approach. This uses auxiliary information for the small areas of interest and has been characterized in the statistical literature as “borrowing strength” from the relationship between the values of the response variables and the auxiliary information (Saei and Chambers, 2005). Small area estimation deals with inference problems for this kind of domains. In these cases, direct estimators might have large sampling errors. Direct estimators can be improved by assuming regression models that link all the sample data by introducing a relation between the variable of interest and a set of explanatory variables (Benavent and Morales, 2016).

Gómez-Rubio et al. (2010) noted that, when some areas are not included in the survey, it is still possible to provide estimates for those areas by relying on the fitted Bayesian models (using in-sample areas) and their spatial correlation to off-sample areas. Area level covariates are still required in all areas to compute the small area estimates. As expected, these estimates are less accurate than in the case with survey data in all areas but, despite the loss of performance, the results are still reasonable and have lower bias and better coverage than traditional synthetic estimates. When data are very sparse, spatial random effects can be incorporated at a regional level, so that larger-scale spatial patterns are modeled. This can help to cope with large amounts of areas with no direct observations and provide reliable results.

Having only a small sample (and possibly an empty sample) in a given area, the only possible solution to the estimation problem is to borrow information from other related data sets (Chandra, 2003). The methods used for small area estimation can be divided accordingly by the related data sources that they employ, whether cross - sectional (from other areas), past data or both. Based on the

level of auxiliary information available methods can also be divided into area level and unit level small area models (Rao, 2003). The spatial (and spatial-temporal) modeling literature centers around two main data types: areal (or lattice) and point-referenced (or geo-statistical). Point-referenced data structures are based on the exact geographical location of an observation being recorded, generally in the form of latitude and longitude co-ordinates (Anderson and Ryan, 2017). This form of data is commonly used for monitoring environmental outcomes, where spatial modeling approaches can be used to characterize the nature of an environmental outcome across the entire study region based on a finite set of monitoring stations. According to Lee et al. (2018) areal databases provide data on a set of K areal units for N consecutive time periods, yielding a rectangular array of $K \times N$ spatial-temporal observations. The motivations for modeling these data are varied, and include estimating the effect of a risk factor on a response see Wakefield (2007) and Lee et al. (2009), identifying clusters of contiguous areal units that exhibit an elevated risk of disease compared with neighboring areas (see Charras-Garrido et al. 2012, and Anderson et al. 2014), and quantifying the level of segregation in a city between two or more different groups (see Lee et al. 2015).

Areal data structures are based on the study region being partitioned into a set of non-overlapping sub-regions known as areal units—for example, a county being divided into a set of postcode areas. Areal data are commonly used in health applications, where confidentiality issues prevent the exact geographical locations of disease cases being recorded. Instead, only the patient's areal unit is recorded, and the data consists of an aggregated count for each individual areal unit (Anderson and Ryan, 2017). However, a common challenge when modeling areal data is that of spatial-temporal auto-correlation, namely that observations from geographically close areal units and temporally close time periods tend to have more similar values than units and time periods that are further apart. Temporal auto-correlation occurs because the data relate to largely the same populations in consecutive time periods, while spatial auto-correlation can arise for a number of

reasons.

The first is unmeasured confounding, which occurs when a spatially patterned risk factor for the response variable is not included in a regression model, and hence its omission induces spatial structure into the residuals. Other causes of spatial auto-correlation include neighborhood effects, where the behaviors of individuals in an areal unit are influenced by individuals in adjacent units, and grouping effects, where groups of people with similar behaviors choose to live close together (Lee et al., 2018). The spatial-temporal models allow for spatial-temporal auto-correlation via random effects, which capture the auto-correlation in the disease data Lawson (2013a). In this study spatial correlation test was conducted to establish the nature of correlation.

2.3 To model cervical cancer cases using Poisson-Gamma and spatial-temporal models.

A normal Poisson model does not account for extra variance, to take into account the extra variance a Poisson-Gamma model is applicable as an alternative. A Poisson-Gamma distribution can be seen as a mixed model, in which gamma distributed random-effects for each area are considered Neyens et al. (2012). The Bayesian inference combines the prior distribution on model parameters and the data likelihood to derive the posterior distribution which summarizes the behavior of the parameters in light of the observed data. According to Lawson (2013a), Bayesian hierarchical models that incorporate time and area effects provide additional insights in terms of the interpretability and similarity based on the neighborhood structure of areas and adjacent times. However, incorporating time and area effects results in increasingly complex model structures which can substantially increase the computational time required to estimate these models. Mapping county level estimates provides greater understanding of the trends and variability in spatial-temporal patterns of less common causes of mortality outcomes not possible by examination of direct national and state estimates Schaible (1996) or

by examination of direct county level estimates.

There are many examples of Bayesian Hierarchical models for small samples and data with excess zeros in the literature. Recently, these methods have gained popularity in epidemiology and public health studies. Khana et al. (2018a) examined existing hierarchical Bayesian spatial-temporal models that account for extra uncertainty, inherent spatial auto-correlation, and the time dependent structure of the data to produce smoothed model based yearly county level Suicide Rates in the software R-INLA. Myer et al. (2017) used a Bayesian Integrated Laplace Approximation and Stochastic Partial Differential Equations (SPDE) method to fit a spatio-temporal model of West Nile Virus (WNV) infection rates in Suffolk County, Long Island, mosquitoes. Oleson and Wikle (2013) used a spatial-temporal hurdle model based on a Gaussian latent process to predict infectious disease outbreak risk via migratory waterfowl vectors. Cnaan et al. (1997) used the INLA approach to implement Bayesian spatial and spatial-temporal zero-inflated models. Other popular models include the Besag-York-Mollié (BYM) model for disease-mapping with extensions for regional data Besag et al. (1991), continuous-indexed Gaussian models (Yue and Wang, 2014, Diggle and Lophaven, 2006). Cervical cancer is an event associated with space and time, consequently counties relative risks estimates can be obtained using Bayesian hierarchical models even in if there are no co-variates.

Among women cervical cancer was the fourth prevalent cancer worldwide (Vidoni et al., 2021). Earlier a study by (Bray et al., 2018) revealed that cervical cancer was the second leading cause of cancer death contributing 10% (3,266 deaths) while breast cancer was third with 7.7% (2,553 deaths). Human Papillomavirus (HPV) infection which is transmitted through direct contact is the cause of almost all cervical cancers (Fontham et al., 2020). According to (Schaafsma et al., 2015) sexually transmitted HPV genotypes, notably HPV16 cause virtually all cervical cancers world-wide if not controlled immunologically or by screening. The control strategies for cervical cancer includes early screening, vaccination against HPV,

treatment of per-cancerous lesions, diagnosis and treatment of invasive cervical cancer and palliative care (WHO, 2020). Cervical cancer screening aides in detection of abnormalities which can be treated and pre-cancers which may progress into actual cancer thus reducing cervical cancer incidences, deaths and morbidity related to treatment(Schaafsma et al., 2015).

2.4 To model the effects of covariates on spatial-temporal distribution of oesophageal and lung cancer cases in Kenya's counties.

The use of Bayesian models in the areas of disease mapping, epidemiology, and small area health applications is well established. Lee (2013) outlined that the set of areal units on which data are recorded can form a regular lattice or differ largely in both shape and size, with examples of the latter including the set of electoral wards or census tracts corresponding to a city or county. In either case such data typically exhibit spatial auto-correlation, with observations from areal units close together tending to have similar values. A proportion of this spatial auto-correlation may be modeled by including known covariate risk factors in a regression model, but it is common for spatial structure to remain in the residuals after accounting for these covariate effects.

Specifically, in a Bayesian spatial-temporal model, the spatially structured and unstructured random effects are used to model the inherent spatial auto-correlation in the data, the correlated and uncorrelated time effects model the time dependent structure of the data, time varying covariates model the extra uncertainty in the data due to measured confounders, and the space-time interaction effects model the residual spatial-temporal variation that are unaccounted for by the county and time random effects to produce reliable model based yearly county level estimates (Khana et al., 2018a). The small-scale geography (e.g. county level) data for a less common cause of mortality outcome, in general, often exhibits strong spatial

auto-correlation. According to Besag et al. (1991) time varying covariates can account for some of the spatial and temporal auto correlation. According to Lawson (2013a), the residual spatial auto-correlation is accounted for by the introduction of spatially structured random effects into the model.

The integrated nested Laplace approximation (INLA) modeling approach provides the ability to use Bayesian inference with a latent Gaussian model fit to large data sets in a short time while using fewer computing resources than commonly used approaches such as the WinBUGS or JAGS Gibbs samplers, which use the more standard and time-consuming Markov chain Monte Carlo algorithms (Rue et al., 2009). The INLA approach is particularly well suited to spatial and temporal models of disease incidence, because they are usually described using latent Gaussian models with a hierarchical Bayesian framework (Schrödle and Held, 2011). Existing Bayesian spatial-temporal modeling strategies can be applied via Integrated Nested Laplace Approximation (INLA) in R to a large number of rare causes of mortality outcomes to enable examination of spatial-temporal variations on smaller geographic scales such as counties (Khana et al., 2018a). Khana et al. (2018a) suggested that future analyses might not require too long of a stretch of data in time in order to compute county level estimates since temporal random effect was found to be an auto-regressive process of order 1 which dampens out after a certain period of time.

The hierarchical Bayes statistical models employ multiple levels of modeling specified in a hierarchical order to estimate the posterior distributions of the model parameters using the Bayes method. The observed data is combined with the multiple sub-level model specifications (prior distributions) and possible covariates to estimate the posterior distribution via Bayes theorem (Khana et al., 2018a).

The modeling of spatially structured random effects via the adjacency matrix of the counties by conditional auto-regressive priors was first proposed by (Besag et al., 1991). To account for potential linear and non-linear trends and extra variation in county level estimates over time, fixed, correlated and uncorrelated

time effects and space time interaction effects are incorporated. Depending on the nature of the data, a variety of latent models such as random walk-1, random walk-2, besag, convolution among others can be implemented via R-INLA software package to model the small area outcome and produce reliable smoothed estimates. Adem et al. (2015) utilized Bayesian Hierarchical Generalized Linear Mixed Models (GHGLMM) with spatial temporal methods to model Tuberculosis cases in Kenya using Small Area Estimation. Lawson and Rotejanaprasert (2014) utilized Bayesian clustering approach to assess the degree of spatial clustering of geocoded data for pediatric brain cancer in Florida. They concluded that there was excess risk in a number of relatively dispersed zip codes across the state but it appeared there was some concentration of high risk in other areas.

Oesophageal cancer is the cancer that forms in tissues lining the oesophagus (the muscular tube through which food passes from the throat to the stomach) (Cancer.Net, 2021). According to study findings by (Schaafsma et al., 2015), (Ferlay et al., 2015) the rate of oesophageal cancer in Kenya was 17.6 per 100,000 which was one of the highest incidence in the Africa continent and was the most common male cancer in Eldoret. Hospital-based studies conducted in Tenwek hospital in western Kenya by Tenge et al. (2009) revealed that male: female ratio of 1.6:1.12 indicating higher incidence rates among males than females. The reasons for the high burden of oesophageal cancers in several parts of Eastern Africa and Southern Africa are not fully understood. Tobacco and alcohol was shown to be clear risk factors in South Africa Pacella-Norman et al. (2002) but obviously do not explain the high rates in East Africa compared with other regions (Korir et al., 2015). Kenya is one of a few countries that lie on Africa's oesophageal cancer corridor, which is a region situated in the geographic area of the Eastern and Western rift-valley and is reported to have the highest oesophageal cancer incidences in Africa (Schaafsma et al., 2015). Therefore a study on the risk factors such as smoking and alcohol use on oesophageal cancer was very appropriate.

In Kenya prostrate, oesophageal and colorectal cancers are the the most preva-

lent among men while breast, cervical and oesophageal cancers are most common among women. Oesophageal cancer contributes 13.2% of cancer mortality which is the highest, cervical is the second contributing 10% of the cancer deaths while breast cancer comes third at 7.7% (Bray et al., 2018). Patel et al. (2013) conducted a study in Moi Teaching and Referral Hospital (MTRH) in Uasin Gishu County where they identified oesophageal cancer as the leading cancer in men. Kenya has a few hospitals which treat oesophageal cancer patients, some of which include Kenyatta National Teaching and Referral Hospital, Moi Teaching Referral Hospital, Tenwek Mission Hospital, Kijabe Mission Hospital, M. P. Shah Hospital/ Cancer Care Kenya. People with oesophageal cancer may experience: difficulty and pain with swallowing, burning in the chest, frequent choking on food and indigestion or heartburn (Cancer.Net, 2021). (Odera et al., 2017) identified alcohol drinking, genetic factors, dietary change/food preparation, and consumption of hot food as the main risk factors for oesophageal cancer in Kenya, they noted that there is a need to investigate the causal relationships between these major risk factors and the development of oesophageal cancer in Kenya. Recent studies on oesophageal cancer has focused on specific regions, therefore mapping its rates, identifying the risk factors as well as locating counties with high rates will help them prioritize control strategies and design ways to modify risk behaviors.

Lung cancer is the cancer that forms in tissues of the lung, usually in the cells lining air passages. According to Bray et al. (2018) worldwide, lung cancer remains the leading cause of cancer incidence and mortality, with 2.1 million new lung cancer cases and 1.8 million deaths predicted in 2018, representing close to 1 in 5 (18.4%) cancer deaths. According to American Cancer Society (2021), the two main types are small cell lung cancer and non-small cell lung cancer. The main risk factor for lung cancer is smoking resulting 80% of deaths, where the percentage might be higher for small cell lung cancer (SCLC). Other risk factors includes: Exposure to asbestos and radon a radioactive gas. Bandera et al. (2001) and Korte et al. (2002) suggested smoking-adjusted association for high alcohol consumption.

Clinical manifestations which may suggest lung cancer include: Respiratory symptoms: coughing, coughing blood, wheezing sound or shortness of breath. Systemic symptoms: weight loss, weakness, fever, or clubbing of the fingernails. Symptoms due to the cancer mass pressing on adjacent structures: chest pain, bone pain, superior vena cava obstruction, or difficulty in swallowing (Kasper et al., 2015). Therefore it is appropriate to conduct the study in Kenya to determine whether smoking and alcohol use are risk factors in cancer patients in Kenya.

CHAPTER THREE

3 METHODOLOGY

3.1 Introduction

This chapter will provide details of the methodologies to be applied in this study.

3.2 Methodologies for the Objectives

3.2.1 To model over-dispersion and conduct spatial correlations tests in order to model cervical, oesophageal and lung cancer cases distribution in Kenya's counties.

Poisson log-normal model

Assessment of presence of over dispersion in the data set was conducted by first computing the residuals from a simple Poisson log-linear model. Poisson log-linear model is a model for n responses, Y_1, \dots, Y_n representing the cancer cases in different counties. There the distribution is

$$Y_i \sim \text{Poisson}(\lambda_i)$$

where λ_i is the rate parameter. The simple Poisson model is stated as follows:

$$\log \lambda_i = \beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip} \quad (3.1)$$

or equivalently

$$\lambda_i = e^{\beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip}} \quad (3.2)$$

The covariates are treated as fixed constants and the model parameters are the $\beta = (\beta_0, \dots, \beta_p)$.

Estimation

The model parameters are estimated using Maximum Likelihood Estimation (MLE) methodology as illustrated below. Y_1, \dots, Y_n are independent Poisson random variables, the likelihood is given by

$$l(\beta_0, \dots, \beta_p) = \prod_{i=1}^n \frac{\lambda_i^{Y_i} e^{-\lambda_i}}{Y_i!} \quad (3.3)$$

where λ_i is defined in terms of β_0, \dots, β_p and the covariates x_{i1}, \dots, x_{ip} via equation (3.1). Setting $x_{i1} = 1$ for all i , the log-likelihood is then

$$\begin{aligned} l(\beta_0, \dots, \beta_p) &= \sum_{i=1}^n Y_i \log \lambda_i - \lambda_i - \log Y_i! \\ &= \sum_{i=1}^n Y_i \left(\sum_{j=0}^p \beta_j x_{ij} \right) - e^{\sum_{j=0}^p \beta_j x_{ij}} - \log Y_i! \end{aligned} \quad (3.4)$$

the MLEs are solutions to system of score equations, for $m = 0, \dots, p$

$$0 = \frac{\partial l}{\partial \beta_m} = \sum_{i=1}^n x_{im} (Y_i - e^{\sum_{j=0}^p \beta_j x_{ij}}) \quad (3.5)$$

These equations may be solved numerically using the Newton-Raphson method. The model assumption of a Poisson distribution for Y_i , is that the variance of Y_i must be equal to its mean which is rather restrictive. This is not usually the case, if observed variance of Y_i is larger than its mean - this is referred to as over dispersion.

Spatial correlation test

Definition: Consider a study area which comprises n counties. Let the observed value of a variate, Y , in county i be y_i . For every pair of counties, i and j , in the study area the drawings which yield y_i and y_j are uncorrelated, then we say that there is no spatial auto-correlation in the county system on Y . Conversely, spatial auto-correlation is said to exist if the drawings are not all pairwise uncorrelated. Measures of spatial auto-correlation describe the degree to which observations (values) at spatial locations (whether they are points or areas), are similar to each

other. So we need two things: observations and locations (Cliff and Ord, 1970).

Compute Moran's I

$$I = \frac{n}{\sum_{i=1}^n (y_i - \bar{y})^2} \frac{\sum_{i=1}^n \sum_{j=1}^n w_{ij} (y_i - \bar{y})(y_j - \bar{y})}{\sum_{i=1}^n \sum_{j=1}^n w_{ij}} \quad (3.6)$$

Where w_{ij} are spatial weights. w_{ij} equals 1 for counties i and j that are deemed neighbors and otherwise 0.

Moran's I, is an inferential statistic, and statistical significance has to be determined before interpreting the index Moran (1950). This is done with a simple hypothesis test, calculating a z-score and its associated p-value.

The null hypothesis for the test is that the data is randomly disbursed. The alternate hypothesis is that the data is more spatially clustered than you would expect by chance alone. Two possible scenarios are: A positive z-value: data is spatially clustered in some way. A negative z-value: data is clustered in a competitive way.

3.2.2 To model cervical cancer cases using Poisson-Gamma and Spatial-temporal models.

Standardized Incidence Ratios

Suppose that the index $s \in (1, 2, \dots, S)$ represents the geographically connected regions. The spatially correlated effects in INLA are introduced by assuming that neighboring regions are more alike than two arbitrary regions. Two regions s and s' are neighbors if they share a common boundary. Moraga (2018), outlines that disease risk estimates in areas can be obtained by computing the Standardized Incidence Ratios. For area i , $i = 1, \dots, n$, the SIR is obtained as the ratio of the observed to the expected disease counts: $SIR_i = \frac{Y_i}{E_i}$. The expected counts represent the total number of disease cases that one would expect if the population of the specific area behaved the way the standard (or regional) population

behaves. The expected counts can be calculated using indirect standardization as $E_i = \sum_{j=1}^m r_j^{(s)} n_j$, where $r_j^{(s)}$ is the disease rate in stratum j of the standard population, and n_j is the population in stratum j of the specific area. The SIR corresponding to area i , SIR_i , indicates whether the area i has more ($SIR_i > 1$), equal ($SIR_i = 1$) or fewer ($SIR_i < 1$) cases observed than expected from the standard population. When applied to mortality data, the ratio is commonly known as the Standardized Mortality Ratio (SMR).

Although in some situations SIRs can give a sense of the disease's spatial variability, very extreme values can occur in areas with small populations owing to the small sample sizes involved. In contrast, disease models are preferred to obtain disease risks estimates because they enable to incorporate covariates and borrow information from neighboring areas to improve local estimates, resulting in the smoothing or shrinking of extreme values based on small sample sizes (Moraga, 2018).

Although SIRs can be useful in some settings, in regions with small populations or rare diseases the expected counts may be very low and SIRs may be misleading and insufficiently reliable for reporting. Therefore, it is preferred to estimate disease risk by using models that enable to borrow information from neighboring areas, and incorporate covariates information resulting in the smoothing or shrinking of extreme values based on small sample sizes (Lawson, 2013a).

Bayesian hierarchical models specifications

Full Bayesian Hierarchical models that is, Poisson-gamma (PG) and Spatial-temporal models with assumption that the response variable was generated by a Poisson process (count data) were formulated to model cervical cancer cases. Fully Bayesian model involve setting prior for group-level parameters as well as hyper-parameters.

(i) Poisson-Gamma model

A Poisson-Gamma (PG) model, with two-level hierarchy Neyens et al. (2012) was considered. Considering observed counts Y_i in area i , are modeled using a Poisson

distribution with mean $E_i\theta_i$, where E_i is the expected counts and θ_i is the relative risk in area i . The relative risk θ_i quantifies whether area i has higher ($\theta_i > 1$) or lower ($\theta_i < 1$) risk than the average risk in the standard population (Moraga, 2018). For example, if $\theta_i = 2$, this means that the risk of area i is two times the average risk in the standard population. The two level hierarchy Poisson-Gamma model can be written as:

$$y_i \sim \text{Poisson}(E_i\theta_i); \quad (3.7)$$

$$\theta_i \sim \text{Gamma}(\alpha, \beta); \quad (3.8)$$

$$\alpha|\nu \sim h_\alpha(\nu)$$

$$\beta|\rho \sim h_\beta(\rho)$$

θ has a $\text{Gamma}(\alpha, \beta)$ distribution at the first level of hierarchy while at the second level α has a hyper prior distribution h_α and β will be h_β .

Due to the conjugacy between the Poisson and gamma distributions, a closed-form posterior distribution can be provided and is given by a gamma distribution with parameters $y_i + \alpha$ and $E_i + \beta$, respectively.

In first level is assumed that the random variable y_i has Poisson distribution or written as $y_i \sim \text{Poisson}(e_i\mu_i\theta_i)$ with probability density function :

$$g(y_i|e_i\mu_i\theta_i) = \frac{e^{-(e_i\mu_i\theta_i)}(e_i\mu_i\theta_i)^{y_i}}{y_i!}, y_i = 0, 1, \dots \quad (3.9)$$

where $\mu_i = \mu(\underline{x}_i, \underline{\beta})$ is regression model, $x_i = (x_{1i}, x_{2i}, \dots, x_{pi})^T$ is a vector of

covariates and $\underline{\beta} = (\beta_1, \beta_2, \dots, \beta_p)^T$ is regression coefficients. In the second stage, it is assumed the parameter due to the conjugacy between the Poisson and gamma distributions, a closed-form posterior distribution can be provided and is given by a gamma distribution with parameters θ_i has Gamma distribution or $\theta_i \sim \text{Gamma}(\alpha, \beta)$ with probability density function (prior on θ) is:

$$k(\theta_i) = \frac{\beta^\alpha}{\Gamma(\alpha)} e^{-\beta\theta_i} \theta_i^{\alpha-1}, \theta_i > 0 \quad (3.10)$$

based on equation above the joint probability density function is obtained as follows:

$$h(y_i|\theta_i) = \frac{e^{-(e_i\mu_i\theta_i)}(e_i\mu_i\theta_i)^{y_i}}{y_i!} \frac{\beta^\alpha}{\Gamma(\alpha)} e^{-\beta\theta_i} \theta_i^{\alpha-1}, y_i = 0, 1, \dots; \theta_i > 0 \quad (3.11)$$

The marginal probability density function is as follows

$$\begin{aligned} m(y_i) &= \int_0^\infty h(y_i, \theta_i) d\theta_i \\ &= \binom{y_i + \alpha - 1}{\alpha - 1} \left(\frac{\alpha}{e_i\mu_i + \alpha} \right)^\alpha \left(1 - \frac{\alpha}{e_i\mu_i + \alpha} \right)^{y_i} \end{aligned} \quad (3.12)$$

The distribution of equation above is Negative-Binomial with mean and variance for y_i respectively are as follows:

$$E(Y_i|\underline{\beta}, \alpha) = e_i\mu_i$$

and

$$\text{Var}(Y_i|\underline{\beta}, \alpha) = e_i\mu_i \left(1 + \frac{e_i\mu_i}{\alpha} \right)$$

The posterior distribution for θ_i is estimated as follows

$$= \frac{(e_i\mu_i + \alpha)^{y_i + \alpha}}{\Gamma(y_i + \alpha)} e^{-(e_i\mu_i + \alpha)} (\theta_i)^{y_i + \alpha - 1}, \theta_i > 0 \quad (3.13)$$

Based on equation above, posterior distribution for θ_i is obtained as : $\theta_i|y_i, \beta, \alpha \sim \text{Gamma}(y_i + \alpha, e_i\mu_i + \alpha)$

Thus, the posterior mean and posterior variance obtained from this Bayes estimate for θ_i are:

$$\widehat{\theta}_i^B(\beta, \alpha) = E_B(\theta_i|y_i, \beta, \alpha) = \frac{(y_i + \alpha)}{(e_i\mu_i + \alpha)}$$

and

$$\text{Var}_B(\theta_i|y_i, \alpha, v) = \frac{(y_i + \alpha)}{(e_i\mu_i + \alpha)^2}$$

According to Miaou and Lord (2003) in statistical literature, Poisson-Gamma model has also been defined as: $\lambda_i = f(\mathbf{X}; \boldsymbol{\beta})\exp(\varepsilon_i) = \mu_i\exp(\varepsilon_i)$ and where, $f(\cdot)$ is a function of the covariates \mathbf{X} , $\boldsymbol{\beta}$ is a vector of coefficients and ε_i is the model error independent of all covariates.

This model only introduces a spatially-unstructured over dispersion factor and that it does not take into account spatial correlation of the data (Neyens et al., 2012). Poisson-Gamma model has been criticized because of the mentioned disadvantage together with the difficulty to include covariates. They has also shown to be inferior to more complex models such as the Conditional Auto-regressive (CAR) convolution models (Lawson et al., 2000). To address the issue of over dispersion, spatial correlation models which include terms for both the over dispersion and the correlated heterogeneity (CH) are utilized.

(ii) Spatial-temporal model

Spatial-temporal model is generated from Generalized Linear Mixed Models (GLMMs). GLMMs is a class of additive structured regression models which have been extensively used to model spatial data in different areas such as in epidemiology, agriculture, demography, economy and image analysis. Their main assumption of the models is that the distribution of the response variable y_i belongs to an exponential family of the form $y_i/\theta_i, \phi_1 \sim p(\cdot)$ defined as,

$$p^{(y_i/\theta_i, \phi_1)} = \exp\left(\frac{y_i\theta_i - b(\theta_i)}{a(\phi_1)} + c(y_i, \phi_1)\right) \quad (3.14)$$

for $i = 1, \dots, n$ observations and θ_i is the scalar canonical parameter, while $a(\phi_1)$ and $c(y_i, \phi_1)$ are known functions. The mean $u_i = E(y_i/\beta f^i(\cdot), \phi_1)$ can be linked to structured additive predictor η_i which accounts for various covariates in an additive way:

$$g(\mu_i) = \eta_i = \beta_0 + \sum_{i=1}^n f_i(\mu_i) + \sum_{k=1}^m \beta_k x_{ki} + \varepsilon_i \quad (3.15)$$

Upon varying the form of the functions $f_i(\cdot)$, this formulation can accommodate a wide range of models, standard and hierarchical regression, spatial and spatial-temporal models or time series. Spatial and spatial temporal models were adopted in our study. Where $f_i(\cdot)$ are unknown functions of the covariates used to model temporal and spatial dependencies and also used to relax the linear relationships of the covariates. The β'_k s represents the linear effects of the covariates x' s and ε'_i s are unstructured terms. Generally the models are described into levels, that is, data likelihood and prior distributions as discussed below;

Level 1: Data likelihood

The region of interest S is the corresponding set of responses $S = S_1, \dots, S_n$, and the vector of known offsets $O = (O_1, \dots, O_n)^T$. The spatial pattern of the response is the matrix of covariates $x = (x_1^T, \dots, x_n^T)^T$ and a series of random effects $\phi = (\phi_1, \dots, \phi_n)$, the latter of which are included to model any spatial auto-correlation that remains in the data after the covariate effects have been accounted for (Lee, 2013). The vector of covariates for areal unit S_i are denoted by $x_i^T = (1; x_{i1}, \dots, x_{ip})$, the first of which corresponds to an intercept term. The general model is an extension of a generalized linear model and is given by

$$Y_i | \mu_i \sim f(y_i | \mu_i, v^2) \text{ for } i = 1, \dots, n \quad (3.16)$$

$$g(\mu_i) = x_i^T \beta + \phi_i + O_i \quad (3.17)$$

Response Y comes from the exponential family of distributions $f(y_i | \mu_i, v^2)$. These can be binomials: Gaussian or Poisson families. Expected value of Y_i is expressed as $E(Y_i) = \mu_i$, where v^2 is an additional scale parameter required when using the Gaussian family. The expected value of the answer is related to the linear predictor through the invertible logic function $g(\cdot)$, like logit (binomial family): identity functions (Gaussian family) or natural logarithm functions (Poisson family). The vector of regression parameters is denoted by $\beta = (\beta_0, \dots, \beta_p)$ and nonlinear covariate effects can be incorporated into the above model by including natural cubic splines or polynomial basis functions in X .

Level 2: Prior distributions

Independence priors

Lee (2013) outlined that for each regression parameter β_j : independent Gaussian priors are given. That is, $\beta_j \sim N(m_j, v_j)$ for $j = 0, \dots, p$: and the software is ($m_j = 0; v_j = 1000$). Gaussian probability scale parameter v^2 is assigned a uniform A distribution: $v^2 \sim U(0, M_v)$: where the diffusion specification $M_v = 1000$ is the default value. Note that a commonly used alternative prior distribution for the variance parameter is Conjugate Inverse Gamma Distribution. However, it is not used here because it is difficult to choose hyper parameters in such a way that it is meaningless for very small values v^2 . Many different random effects models can be implemented. The simplest is independence prior

$$\theta \sim N(0, \delta^2)$$

$$\delta^2 \sim U(0, M_\sigma)$$

; where θ_k replaces ϕ_k in the data likelihood. The variance parameter is assigned a uniform prior on the interval $(0, M_\sigma)$, where before the default value $M_\sigma = 1000$. This specification is appropriate when the covariates included in the model (3.11)

remove all spatial structure of the response: leaving random effects to account for possible effects of scattering (binomial model and Poisson models). However, in most data sets there is likely to be residual spatial auto-correlation, in which case one of the global priors described below is required.

Global Conditional Auto-regressive (CAR) priors

Four different conditionally auto-regressive priors (CAR) are commonly used for modeling spatial auto-correlation in the statistics literature, the intrinsic and Besag,-York-Mollié (BYM) models (Besag et al. 1991) as well as the alternatives developed by Leroux et al. (2000). Each model is a special case of a Gaussian Markov random field (GMRF), and can be written in the general form $\phi \sim N(0, \tau^2 Q^{-1})$, where Q is a precision matrix that may be singular (intrinsic model).

This matrix controls the spatial auto-correlation structure of the random effects, and is based on a non-negative symmetric $n \times n$ neighborhood or weight matrix W . A binary specification based on geographical contiguity is most commonly used, where $w_{ij} = 1$ if areal units ($S_i; S_j$) share a common border (denoted $i \sim j$), and is zero otherwise. This specification forces (ϕ_i, ϕ_j) relating to geographically adjacent areas (that is $w_{ij} = 1$) to be correlated, whereas random effects relating to non-contiguous areal units are conditionally independent given the values of the remaining random effects.

CAR priors are commonly specified as a set of n univariate full conditional distributions

$f(\phi_i | \phi_{-i})$ for $k = 1, \dots, n$ (where $\phi_i = (\phi_1, \dots, \phi_{i-1}, \phi_{i+1}, \dots, \phi_n)$), rather than via the multivariate specification described above. The first CAR prior to be proposed was the intrinsic model (Besag et al., 1991), which is given by

$$\phi_i | \phi_{-i} \sim N\left(\frac{\sum_{j=1}^n w_{ij} \phi_j}{\sum_{j=1}^n w_{ij}}, \frac{\tau^2}{\sum_{j=1}^n w_{ij}}\right) \quad (3.18)$$

The conditional expectation is the average of the random effects in neighboring

areas, while the conditional variance is inversely proportional to the number of neighbors.

The latter is appropriate because if the random effects are spatially correlated, then the more neighbors an area has the more information there is from its neighbors about the value of its random effect.

In common with the other variance parameters, τ^2 is assigned a uniform prior on the interval $(0; M_\tau)$, with the default value being $M_\tau = 1000$. The limitation with this model is that it can only represent strong spatial auto-correlation, and is well known to produce random effects that are overly smooth. Therefore, the same authors proposed an extension to allow for both weak and strong spatial auto-correlation, by replacing ϕ_i in (3.13) with $\theta_i + \phi_i$, which are respectively defined by (3.14) and (3.15). This model is known as the BYM or convolution model, and is the most commonly used CAR model in practice. However, it requires two random effects to be estimated for each data point, whereas only their sum is identifiable from the data. Therefore, Leroux et al. (2000) and Stern and Cressie (1999) proposed alternative CAR priors for modeling varying strengths of spatial auto-correlation, using only a single set of random effects. The model by Leroux et al. (2000) is given by

$$\phi_i | \phi_{-i} \sim N\left(\frac{\rho \sum_{j=1}^n w_{ij} \phi_j}{\rho \sum_{j=1}^n w_{ij} + 1 - \rho}, \frac{\tau^2}{\rho \sum_{j=1}^n w_{ij} + 1 - \rho}\right) \quad (3.19)$$

while the proposal of Stern and Cressie (1999) is

$$\phi_i | \phi_{-i} \sim N\left(\frac{\rho \sum_{j=1}^n w_{ij} \phi_j}{\rho \sum_{j=1}^n w_{ij}}, \frac{\tau^2}{\rho \sum_{j=1}^n w_{ij}}\right) \quad (3.20)$$

In both cases ρ is a spatial auto-correlation parameter, with $\rho = 0$ corresponding to independence, while $\rho = 1$ corresponds to strong spatial auto-correlation. A uniform prior on the unit interval is specified for ρ , that is $\rho \sim U(0, 1)$, while the usual uniform prior on the interval $(0, M_\tau)$ is adopted for τ^2 . In both cases when $\rho = 1$ the intrinsic model proposed by Besag et al. (1991) is obtained, while

when $\rho = 0$ the only difference is the denominator in the conditional variance.

The approaches for Bayesian inference on latent Gaussian models are Markov chain Monte Carlo (MCMC) sampling and integrated nested Laplace approximation (INLA). The high dimensionality of the latent field Θ and the strong correlation within Θ and between Θ and Φ especially when the numbers of observations are many leads to problems in convergence and computation time. INLA developed by Rue et al. (2009) bypasses MCMC entirely by basing inferences on closed form approximations making it computation efficient compared to MCMC since it does not use iterative computation techniques like MCMC. In our spatial-temporal model INLA estimation was utilized.

Spatial-temporal model for this study

Spatial-temporal models for spatial data enables borrowing of information from neighboring areas, and incorporate covariates information resulting in the smoothing or shrinking of extreme values based on small sample sizes (Gelfand et al. 2010; Davis et al. 2009). The classical parametric formulation was introduced by Bernardinelli et al. (1995), and assume that the linear predictor can be written as:

$$\eta_{it} = \beta_0 + v_i + \nu_i + \beta t \quad (3.21)$$

This formulation includes spatial structured effects v_i , unstructured components ν_i and a main linear trend β , which represents the global time effect. The parameters estimated by INLA are $\theta = \{\beta_0, \beta, v, \nu\}$ and the hyper-parameters are represented by $\psi = \{\tau_v, \tau_\nu, \tau_\delta\}$.

Given y_i 's are the observed cervical cancer cases while E_i 's are the expected population per county, y_i 's were assumed to be generated through a Poisson process as shown in equation (3.20). y_i 's are modeled using a Poisson distribution with mean $E_i\theta_i$, where E_i is the expected counts and θ_i is the relative risk in area i . The logarithm of the relative risk θ_i is expressed as the sum of an intercept that models the overall disease risk level, and random effects to account for extra-Poisson

variability.

$$y_i \sim Po(E_i\theta_i), i = 1, \dots, n, \quad (3.22)$$

$$\log(\theta_i) = \eta_{it} = \beta_0 + v_i + \nu_i + \beta t \quad (3.23)$$

In equation (3.23):

1. \log is a monotonic link function for count data. The logarithm of the relative risk θ_i is expressed as the sum of an intercept that models the overall disease risk level, and random effects to account for extra-Poisson variability.
2. β_0 represents the overall risk in the region of study
3. the structured spatial effects, v_i , were estimated at county level in which a household was located and Kenya counties' boundaries were used to compute the neighborhood information. The spatial effects by county to account for strong spatial auto-correlation, and was modeled via normal conditionally auto-regressive priors (CAR) Besag et al. (1991) where weights were assigned to each county according to adjacency; neighboring counties received a weight of one while non-neighboring counties received a weight of zero. Specifically, for $i = 1, \dots, m$, counties and $j = 1, \dots, T$, years; $\phi_i | \phi_{-i}, \tau_v \sim N\left(\frac{\sum_{i=1}^n w_{ij} \phi_i}{\sum_{i=1}^n w_{ij}}, \frac{\tau^2}{\sum_{j=1}^n w_{ij}}\right), \frac{1}{n_{\delta_i} \tau_u}, i \neq j$ where, τ_v is the conditional precision of spatial random effects and δ_i is the neighborhood of the i^{th} region, n_{δ_i} is the number of neighbours, $\sum_{j=1}^m w_{ij}$, and the spatial weight, w_{ij} equals 1 for counties i and j that are deemed neighbors and otherwise 0. According to Bivand (2019) each county has at least one neighbor, and the number of neighbors is determined empirically based on the spatial distribution of the counties. The conditional precision of the spatial random effect was assigned $\tau_v \sim \text{Gamma}(1, 0.001)$ prior.
4. unstructured random effects ν_i by county models residual spatial variation not dealt with by our spatial random effects and was assigned a Normal

prior, $\nu_i \sim N(0, \frac{1}{\tau_v})$, with precision, τ_v . The conditional precision of the unstructured random effect was assigned $\tau_v \sim \text{Gamma}(1, 0.001)$ prior.

The precision's for the intercept, fixed effects and the random effects were assigned priors that are default in R-INLA. INLA assigns $\log(\text{precisions}) \sim \text{log-gamma}(1, 0.001)$ priors (Bivand et al., 2015), (Rue et al., 2009). Various models can be fitted by changing parametrisation of equation (3.23). The relative risks of each area can be obtained without the covariates.

Parameters in spatial-temporal models in equation (3.23) are estimated by assigning Gaussian priors to $\beta'_i s$, $f_i(\cdot)$'s and $\varepsilon'_i s$. This can be represented as $\Theta = (\beta'_k s, f'_i s, \dots)$ where Θ is unobserved multivariate Gaussian random variable, whose density $\pi(\Theta/\phi)$ is controlled by a vector of hyper-parameters Φ (Rue and Martino, 2007). The latent Gaussian field Θ is assumed to have a Gaussian distribution with zero mean and variance covariance matrix $Q(\phi_2)$; with vector of hyper-parameters defined as $\Phi = (\phi_1, \phi_2)$ which are not necessarily Gaussian (Martins et al., 2013a). Latent Gaussian model is composed of three elements namely; the likelihood of the data $\pi(y/\Theta)$, the Gaussian density of the random vector Θ , $\pi(\Theta/\Phi)$ and the prior distribution of the parameter vector $\pi(\Phi)$.

The posterior is therefore defined as

$$\pi(\Theta, \Phi/y) \propto \pi(\Theta) \pi(\Theta/\Phi) \prod_{i=1} \pi(y_i/x_i, \Phi) \quad (3.24)$$

The main inferential interest involves computing the posterior marginals for x_i and posterior marginals for Φ or some Φ_j .

Estimation of parameters in latent Gaussian models

Integrated nested Laplace approximation (INLA) methodology is an appropriate inference based method for approximating the posterior marginals of the latent Gaussian field $\pi(x_i/y)$, $i = 1, \dots, n$ in three steps.

The posterior marginals of the latent effects Θ are written as

$$\pi(x_i/y) = \int \pi(x_i/\Phi, y)\pi(\Phi/y)d\Phi \quad (3.25)$$

$$\pi(\Phi_i/y) = \int \pi(\Phi/y)d\Phi_{-j} \quad (3.26)$$

The posterior marginals $\tilde{\pi}(x_i/y)$ and $\tilde{\pi}(\Phi_i/y)$ can be approximated using the Laplace approximation. The first approximation to $\pi(\Phi/y)$ using Gaussian distributions is constructed as follows

$$\pi(\Phi/y) = \frac{\pi(\Theta, \Phi, y)}{\tilde{\pi}_G(\Theta/\Phi, y)} \Big|_{\Theta=\Theta^*(\Phi)} \quad (3.27)$$

where $\tilde{\pi}_G(\Theta/\Phi, y)$ is a Gaussian approximation to the full conditional of Θ and $\Theta^*(\Phi)$ is the mode of the full conditional for Θ , for a given value of Φ . It involves locating the mode of $\tilde{\pi}(\Phi/y)$ which is used to integrate out the uncertainty with respect to Φ when approximating the posterior marginal of x_i .

The posterior marginals of the latent field are supposed to start from $\tilde{\pi}_G(x_i/\Phi, y)$ and approximate the density of $x_i/\Phi, y$ with the Gaussian marginal derived from $\tilde{\pi}_G(\Theta/\Phi, y)$ i.e

$$\tilde{\pi}(x_i/\Phi, y) = N(x_i; (\Phi), \delta_{ii}^2(\Phi)) \quad (3.28)$$

The marginals of the interest can be computed using numerical integration over a multidimensional grid of values of Φ

$$\tilde{\pi}(x_i/y) = \sum_k \tilde{\pi}(x_i/\Phi_k, y)\tilde{\pi}(\Phi_k/y)\Delta_k \quad (3.29)$$

where the sum is over the values of Φ with area weights Δ_k , which would be equal to 1 if all support points would be equi-distantly chosen (Rue and Martino, 2007).

The first step in INLA computation involves approximating the posterior marginal of Φ by using Laplace approximation in equation (3.30).

The second step involves computing the Laplace approximation of $\tilde{\pi}(x_i/\Phi, y)$ for

selected values of Φ which improves the Gaussian approximation in equation (3.29)

$$\tilde{\pi}_{LA}(x_i/\Phi, y) \propto \frac{\pi(\Theta, \Phi, y)}{\tilde{\pi}_{GG}(\Theta_{-i}/x_i, \Phi, y)} \Big|_{\Theta_{-i} = \Theta_{-i}^*(x_i, \Theta)} \quad (3.30)$$

where $\tilde{\pi}_{GG}(\Theta_{-i}/x_i, \Phi, y)$ is a Gaussian approximation to $\Theta_{-i}/x_i, \Phi, y$ around its mode $\Theta_{-i}(x_i, \Phi)$. An improved version of $\tilde{\pi}_{LA}(x_i/\Phi, y)$ known as Simplified Laplace approximation was developed by (Rue et al., 2009). It involves a series of expansion of $\tilde{\pi}_{LA}(x_i/\Phi, y)$ around $x_i = \mu_i(\Phi)$ which corrects for skewness and location and it is also less computationally expensive (Rue et al., 2009). The third step involves combining steps 1 and 2 using numerical integration in equation 3.29 (Rue et al., 2009) .

3.2.3 To model the effect of covariates on spatial distribution of oesophageal and lung cancer cases in Kenya's counties

To achieve this objective a Generalized Linear Mixed Model assuming a Poisson distribution with spatial (structured and unstructured), interaction and temporal random effects was used to characterize the relationships between cancer cases and covariates. In this model the response variable was generated by a Poisson process:

The Poisson regression model is

$$y_i \sim \text{Poisson}(\mu_i)$$

$$\mu_i = \exp(X_i\beta + \text{offset}_i)$$

where y_i is the observed cancer cases , X_i are the covariates for the i^{th} observation and offset term represented the i^{th} population. The generalized linear mixed model used to describe the cancer cases y_i was of the form:

$$g(\mu_i) = \beta_0 + \sum_j \beta_j X_{ij} + f_{str}(S_i) + f_{unstr}(S_i) + f_{trend}(time) \quad (3.31)$$

Where f_{str} and f_{unstr} are structured and unstructured spatial effects of the counties.

1. $g(\cdot)$ is a monotonic link function, in this case the *log*.
2. β_0 an overall intercept term. The intercept, β_0 was assigned a flat prior.
3. β_j represents vector of regression parameters, the parameter vector of the covariates X_{ij} (Alcohol use and Smoking). $\beta_j X_{ij}$: is the i^{th} row and j^{th} column of the covariates matrix \mathbf{X} and β is a vector of regression parameters. The β for fixed effects ($\beta_j X_{ij}$) was assigned Normal priors $\beta \sim N(0, 100)$. In our model β_j 's are the coefficients of the proportion of smokers and alcohol users of covariates.
4. the spatial effects, $f_{str}(S_i)$ modeled via normal conditionally auto-regressive priors (CAR) (Besag et al., 1991). The conditional precision of the spatial random effect was assigned $\tau_u \sim \text{Gamma}(1, 0.001)$ prior.
5. unstructured random effects $f_{unstr}(S_i)$ by county, to model residual spatial variation not dealt with by our spatial random effects was assigned a Normal prior, $f_{unstr} \sim N(0, \frac{1}{\tau_v})$, with precision $\tau_v \sim \text{Gamma}(1, 0.001)$ prior.
6. Correlated random time effects $f_{trend}(time)$, to account for time dependence, was modeled via first order random walk with precision; conditional precision was assigned $\tau_\delta \sim \text{Gamma}(1, 0.001)$ prior.

The parameters in this model were estimated using Integrated Nested Laplace Approximation (INLA) methodology. Parameters of interest were used to calculate the relative risks which were mapped into the different geographical areas for the different years. The relative risk was presented as μ_i : ($\mu_i > 1$) indicated higher disease risk, ($\mu_i < 1$) lower risk while ($\mu_i = 1$) no risk.

3.3 Model selection criteria

3.3.1 Deviance Information Criteria

The Deviance Information Criterion (DIC) Spiegelhalter et al. (2002) is a popular information criterion designed for hierarchical models, and (in most cases) is well defined for improper priors. Its main application is Bayesian model selection, but it also provides a notion of the effective number of parameters. The deviance is $D(x, \theta) = -2 \sum_{i \in I} \log \pi(y_i | x_i, \theta) + \text{constant}$. The effective number of parameters is the mean of the deviance minus the deviance of the mean. The mean of the deviance can be computed in two steps: first, compute the conditional mean conditioned on θ using univariate numerical integration for each $i \in I$; second, integrate out θ with respect to $\pi(\theta | y)$. The deviance of the mean requires the posterior mean of each $x_i, i \in I$, which is computed from the posterior marginals of x_i 's. Regarding the hyper-parameters, we prefer to use the posterior mode θ^* , as the posterior marginal for θ can be severely skewed.

A set of models following the above general space time modeling approach was explored to determine the contribution of different components, namely, the correlated and uncorrelated random time effects, spatially structured and unstructured random effects, space time interaction term and the different covariates to examine spatio-temporal variation in county level cancer rates. DIC is based on the deviance of the model penalized for model complexity and its interpretation is similar to the Akaike Information Criterion (AIC), with models having smaller DIC being preferred.

3.4 Ethical Approval

The cancer registry had ethics approval from Scientific and Ethics Research (SERU) unit from KEMRI and the Ministry of Health which allowed it to document cancer occurrence in the country. The data was collected by active case finding methods whereby staff from the registry visited various health facilities that diag-

nose and treat cancer within the defined populations of coverage retrospectively and prospectively.

3.5 Sample and Sampling Technique

The area of the study was the 47 counties in Kenya. In this study, count data was considered thus the available count for each county was appropriate to be applied in developing the models. Consequently, there was possibility of non sampling error, to deal with it, the spatial temporal models applied distinguished between uncertainty in the quantity of interest and sampling and non sampling variance Foreman et al. (2012).

3.6 Data Analysis

Spatial-temporal models arise when data are collected across time as well as space and has at least one spatial and one temporal property. An event in a spatial-temporal data set describes a spatial and temporal phenomenon that exists at a certain time t and location x . The data was obtained from a 2-year retrospective County based surveillance study of all reported cancer cases in ten counties namely Bomet, Embu, Kakamega, Kiambu, Machakos, Meru, Mombasa, Nairobi, Nakuru, and Nyeri County conducted in 2015 and 2016 by National Cancer Registry in Kenya. Period which complete data was available for the ten counties. Data sheet had the following variables: sex, age at the time of diagnosis, place of residence, smoking status, alcohol drinking status, and cancer diagnosis that was based on the international classification of disease for oncology (ICD-O). Cervical cancer cases were 1064, oesophageal cancer cases 1599 while lung cancer cases were 256. The data in this study was analyzed using (INLA) and hglm packages in R-programming statistical software. The packages contains functions for fitting Bayesian Hierarchical Generalized Linear Mixed Models (BHGLMMs) and Poisson-Gamma models respectively.

CHAPTER FOUR

4 RESULTS AND DISCUSSION

4.1 Introduction

Data in this study was analyzed using Spatial-temporal model R-packages. The packages contains functions for Bayesian Hierarchical Generalized Linear Mixed Model (BHGLMM). The study findings were presented based on each specific objective.

4.2 To model over-dispersion and conduct spatial correlations tests in order to model cervical, oesophageal and lung cancer cases distribution in Kenya's counties.

4.2.1 Assessing the presence of over-dispersion and spatial correlation for cervical cancer cases

The spatial-temporal models allow for spatial-temporal auto-correlation via random effects, which capture the auto-correlation in the disease data after the effects of the known covariates have been accounted for. Therefore, assessment of presence of over dispersion in the data set was conducted by first computing the residuals from a simple Poisson log-linear model. A histogram was obtained shown in Figure 4.1 in order to check distribution of the data.

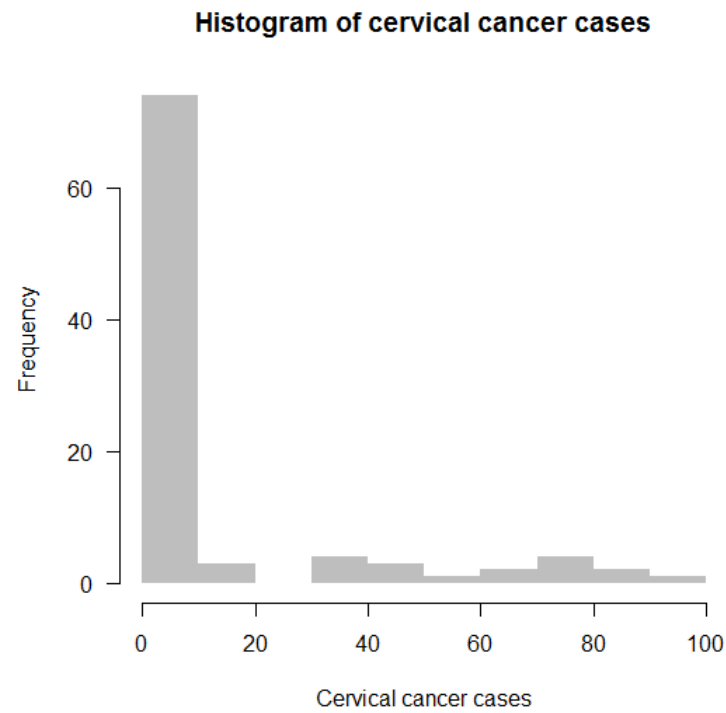


Figure 4.1: Histogram of cervical cancer cases.

Source: Author, 2021

The over dispersion was checked by fitting the Poisson log normal model in equation (3.1), the rule of thumb is that the deviance divided by the degrees of freedom should be equal to 1.

Table 1: Poisson log normal model for cervical cancer cases

Coefficients	df	Estimate	Std. Error	z value	p-value
Intercept		8.093e-11	2.501e-02	0	1
Akaike In-formation Criterion (AIC)	4007.4				
Null deviance (value)	3885.3	93			
Residual deviance (value)	3885.3	93			

In our case, the residual deviance 38885.3 with 93 degrees of freedom. The ratio of deviance to df should be 1, but it in our case $3885.3/93 = 47.77$, indicating over dispersion. A more formal over dispersion test was conducted.

Table 2: Over dispersion test for cervical cancer cases

Coefficients	Value	z value	p-value
Intercept		4.15	1.662e-05
Dispersion parameter	31.22017		

The dispersion parameter is 31.2017, therefore, there is substantial over dispersion, meaning that a simple Poisson regression was not appropriate to model the cervical cancer data.

To quantify the presence of spatial auto-correlation in the residuals from this model we computed Moran's I statistic Moran (1950) using equation (3.6), and conducted a permutation test for each year of data separately. The permutation test has the

null hypothesis of no spatial auto-correlation “the cervical cancer values are randomly distributed across counties following a completely random process” and an alternative hypothesis of positive spatial auto-correlation. The estimated Moran’s I statistic was 0.0399 and the p-value was $0.2104 > 0.05$, suggesting there was no unexplained spatial auto-correlation in the residuals. Therefore the cervical cancer cases were not spatially clustered. Moran’s I statistic is significant and positive when the observed values of locations within a certain distance (d) tend to be similar, negative when they tend to be dissimilar, and approximately zero when the observed values are arranged randomly and independently over space.

Moran I Statistics

Table 3: Moran I Statistics for cervical cancer cases

Statistic	Observed rank	p-value
0.0398	7850	0.2151

4.2.2 Assessing the over dispersion and spatial correlation of oesophageal cancer cases

A simple Poisson regression model was fitted as shown in the output below.

Table 4: Poisson log normal model for oesophageal cancer cases

Coefficients	value	df	Estimate	Std. Error	z value	p-value
Intercept			8.093e-	2.501e-02	0	1
11						
Akaike In-formation Criterion (AIC)	4007.4					
Null deviance	3885.3	93				
Residual deviance	3885.3	93				

In our case, the residual deviance 38885.3 for 93 degrees of freedom. The ratio of deviance to df should be 1, but it is $3885.3/93 = 47.77$, clearly there is over dispersion. A over dispersion test was conducted as shown in output below.

Table 5: Over dispersion test for oesophageal cancer cases

	Value	z value	p-value
Intercept		3.9725	3.555e-05
Dispersion parameter	49.409		

The dispersion parameter is 49.2409 indicating substantial over dispersion, leading to fitting other model which takes care of over dispersion instead of a simple Poisson model.

Spatial auto correlation in our data was assessed by computing the residuals from a simple Poisson log-linear model.

The estimated Moran's I statistic was 0.0399 and the p-value is $0.4155 > 0.05$, indicating there was no spatial auto correlation for oesophageal cancer.

Table 6: Moran I Statistics for oesophageal cancer cases

Statistic	Observed rank	p-value
-0.011379	5846	0.4155

4.2.3 Assessing the over dispersion and spatial correlation of lung cancer data

A simple Poisson-model was fitted as shown in the results below to check over dispersion.

Table 7: Poisson log normal model for lung cancer cases

	Value	df	Estimate	Std. Error	z	p-value
Intercept			3.356e-11	8.084e-02	0	1
Akaike In-formation Criterion (AIC)	484.23					
Null deviance	420.81	93				
Residual deviance	420.81	93				

The residual deviance 420.81 for 93 degrees of freedom. The ratio of deviance to df should be 1, but it is $420.81/93 = 4.52$, indicating over dispersion. An over dispersion test was conducted.

Table 8: Over dispersion test for lung cancer cases

	Value	z value	p-value
Intercept		2.2419	0.0124
Dispersion parameter	6.134		

The dispersion parameter is 6.1342 indicating substantial over dispersion, therefore a simple Poisson regression was not appropriate to model the lung cancer data. Model taking care of over dispersion was fitted in the sections below.

The Moran's I statistic was -0.0133 and the p-value is 0.4120 > 0.05, suggesting there was no unexplained spatial auto correlation in the residuals. Therefore the lung cancer cases were not spatially clustered.

Table 9: Moran I Statistics for oesophageal cancer cases

Statistic	Observed rank	p-value
-0.013258	5881	0.412

4.3 To model cervical cancer cases using Poisson-Gamma and Spatial-temporal models.

4.3.1 Standard Incidence Ratio (SIR) map for cervical cancer cases

To display the distribution of notified cases, Standard Incidence Ratio (SIR), relative risks and cervical cancer distribution in Kenya, spatial temporal maps were produced. This section display the distribution of notified cases, Standard Incidence Ratio (SIR) map, for all counties with notified cervical cancer cases over two years (2015-2016). SIR_i indicates whether area has higher ($SIR_i > 1$), equal ($SIR_i = 1$) or lower ($SIR_i < 1$) risk than expected from the standard population.

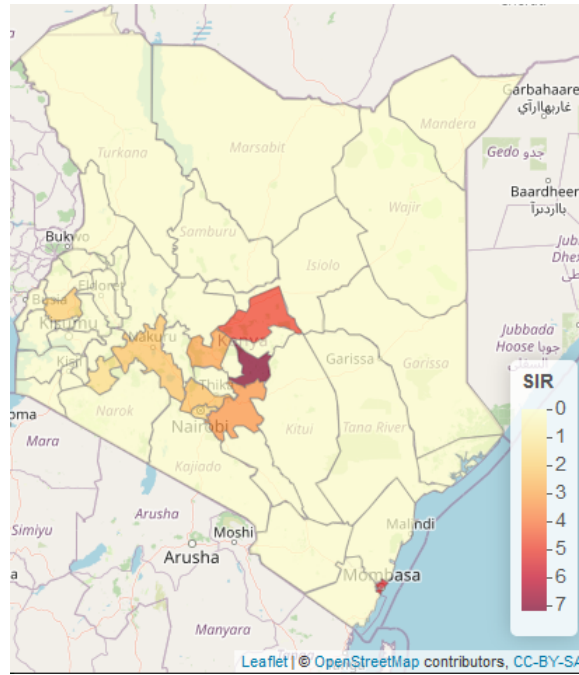


Figure 4.2: **Standardized Incidence Ratio (SIR).**

Source: Author, 2021

Clearly, from Figure 4.2 in most counties there was great risk of cervical cancer cases than expected from the standard population since all counties where data was available had a SIR value greater than 1. Bomet Standardized Incidence Rate value was 1.59, Embu =7.13, Kakamega =2.02, Kiambu =2.42, Machakos =3.44, Meru=4.82, Mombasa =5.51, Nairobi=1.66, Nakuru =2.26, Nyeri =3.07. The deep purple areas exhibited elevated risks ($SIR > 1$) while the light shaded areas were low risk ($SIR < 1$).

4.3.2 Poisson-gamma model

A generalized linear mixed-effects model with both fixed effects and random effects was explored for this study. A Poisson-Gamma model that take care of over dispersion model and zero inflated variables was fitted from equation (3.7) and (3.8).The dispersion parameter for random effect was 1.8692 which was close to 1 and the Akaike Information Criterion (AIC) was 262.9605.

Table 10: Poisson-Gamma model: Summary of the fixed effects estimates

Coefficients	Estimate	Std. Error	z value	p-value
Intercept	-0.3178	0.3731	-0.852	0.399
as.factor(NAME_1) Baringo	0.0144	2.5729		
as.factor(NAME_1) Bomet	2.1385	0.4115		
as.factor(NAME_1) Bungoma	0.0058	2.5734		
...				

The results in Table 11 indicates that, the highest burden of cervical cancer cases was in Embu, Mombasa, Meru, Machakos and Nyeri counties respectively.

Table 11: Relative risks for cervical cancer Poisson-Gamma model

County	Relative Risk 2015/2016
Bomet	2.14
Embu	9.89
Kakamega	2.78
Kiambu	3.40
Machakos	4.76
Meru	6.48
Mombasa	7.41
Nairobi	2.28
Nakuru	2.19
Nyeri	4.28

4.3.3 Spatial-temporal models for cervical cancer cases

Although Standardized Incidence Ratio(SIRs) can be useful in some settings, in regions with small populations or rare diseases the expected counts may be very

low and SIRs may be misleading and insufficiently reliable for reporting. Therefore, it is preferred to estimate disease risk by using models that enable to borrow information from neighboring areas, and incorporate covariates information resulting in the smoothing or shrinking of extreme values based on small sample sizes (Gelfand et al. 2010; Davis et al. 2009).

The first model was fitted based on equation (3.23) in R-INLA. The second model was obtained where the assumption of linearity was released using a dynamic non parametric formulation for the linear predictor defined in equation (3.23). The model contained structured spatial effects, unstructured spatial effects while the temporally structured effect were modelled dynamically (e.g. using a random walk) through a neighboring structure.

A third model was expanding model (3.23) which allowed for an interaction between space and time, which explained differences in the time trend of cancer cases for different areas was implemented.

Table 12 presents the DIC components for the three models: the third model with the dynamic parameterization of the time trend and the space-time interaction had a smaller DIC suggesting that, despite the added complexity, this model had a more appropriate fit to the data. For this reason this was the plausible model and the relative risks were obtained from the model.

Table 12: Deviance Information Criterion (DIC) for the three Spatial-temporal models

Model	\bar{D}	p_D	DIC
Model 1	167.4847	49.6866	217.1713
Model 2	175.0441	30.7786	205.8228
Model 3	174.5819	30.7797	205.3616

The key interest in this analysis of spatial-temporal models is the effects of the disease risk, which for Poisson models are typically presented as relative risks. The posterior relative risk distributions greater than 1 indicates an elevated risk of the disease.

The results in Table 13 were consistent with Poisson-Gamma model results above since Embu county had the highest relative risk of cervical cancer followed by Mombasa, Meru, Machakos and Nyeri, Kiambu, Kakamega, Nakuru, Nairobi and Bomet respectively.

Table 13: Relative risks for cervical cancer spatial temporal model

County	Relative Risks
	2015/2016
Bomet	1.53
Embu	7.92
Kakamega	2.06
Kiambu	2.80
Machakos	3.48
Meru	4.43
Mombasa	5.12
Nairobi	1.63
Nakuru	1.95
Nyeri	3.33

In Figure 4.3 the light yellow coloured areas indicated low risk areas while the purple area indicated an elevated cervical cancer risk, while in Figure 4.4 the purple coloured areas indicated posterior probabilities above 0.8.

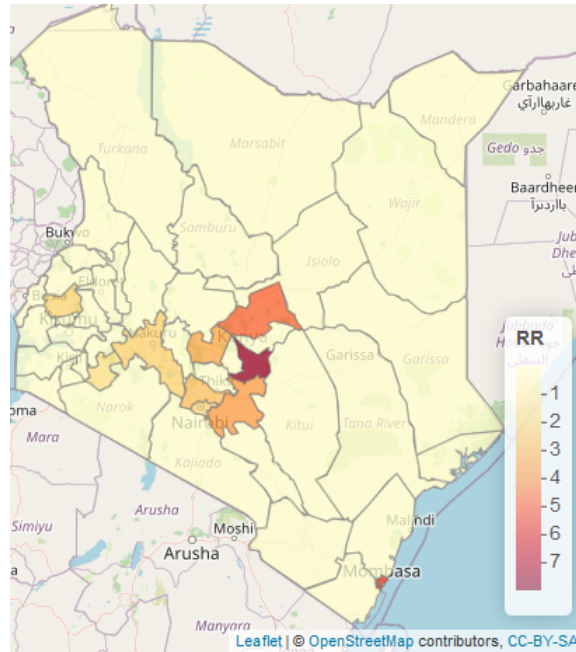


Figure 4.3: Distribution of the county specific relative risks of cervical cancer in the disease mapping model.

Source: Author, 2021

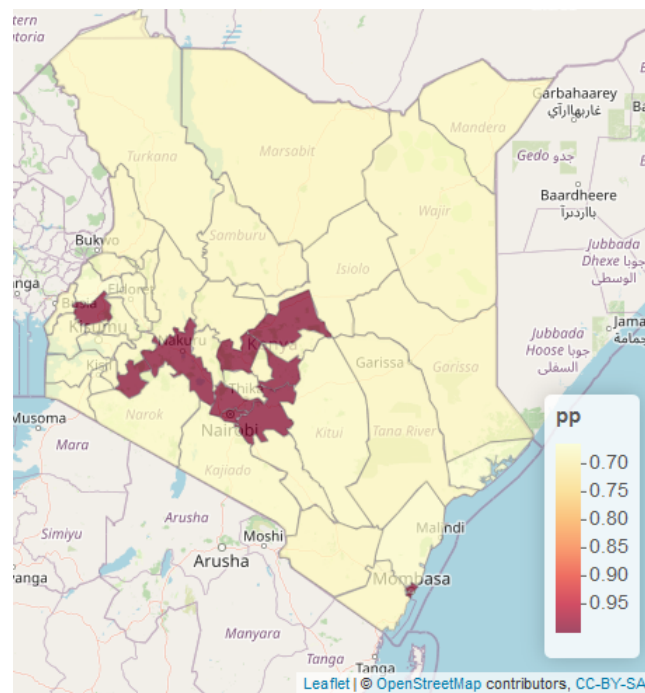


Figure 4.4: Map of the uncertainty for the spatial effect $\zeta_i: p(\zeta_i > 1|y)$.

Source: Author, 2021

4.4 To model the effects of covariates on spatial-temporal distribution of oesophageal and lung cancer cases in Kenya's counties.

4.4.1 Descriptive statistics for oesophageal cancer

Table 14: Distribution of oesophageal cancer by gender in 2015

Gender	Count of Gender	Percentage
Female	349	44.52
Male	435	55.48
Grand Total	784	100

According to data in Table 14, 435 (55.48%) of oesophageal cancer cases were male while 349 (44.52%) of the cases were female.

Table 15: Distribution of oesophageal cancer by gender in 2016

Gender	Count of Gender	Percentage
Female	289	35.46
Male	526	64.54
Grand Total	815	100

In 2016 as shown in Table 5, 526 (64.54%) of oesophageal cancer cases were male while 289(35.46%) of the cases were female.

4.4.2 Standardized Incidence Ratio (SIR) of oesophageal cancer

Clearly in most counties there was greater risk of oesophageal cancer cases than expected from the standard population since all counties where data was available had a SIR value greater than 1 except in Kiambu as shown in Table 16.

Table 16: Oesophageal cancer Standardized Incidence Ratios (SIR)

County	SIR 2015/2016
Bomet	10.09
Embu	4.25
Kakamega	1.91
Kiambu	0.87
Machakos	1.39
Meru	4.22
Mombasa	1.09
Nairobi	2.4
Nakuru	3.08
Nyeri	6.34

Standard Incidence Ratios(SIR) map was generated as shown in Figure 4.5 below

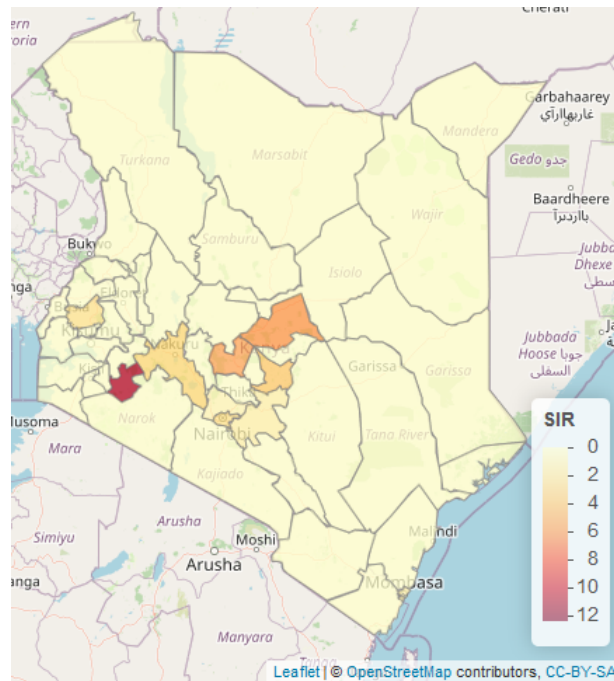


Figure 4.5: Standardized Incidence Rates (SIR) for oesophageal cancer.

Source: Author, 2021

4.4.3 Spatial-temporal models for oesophageal cancer

Although SIRs can be useful in some settings, in regions with small populations or rare diseases the expected counts may be very low and SIRs may be misleading and insufficiently reliable for reporting. Therefore, it is preferred to estimate disease risk by using models that incorporates covariates and borrows information from neighboring areas (Gelfand et al., 2010).

Four models were fitted based on equation (3.31), thereafter a most plausible model was selected based on the smallest value of Deviance Information Criterion (DIC).

Models where smoking was the covariate

Model 1: With structured, unstructured spatial effect, trend effects.

Model 2: With structured spatial effect, structured trend effect, global time effect and a covariate.

Model 3: With structured, unstructured spatial effects, structured trend effects and a covariate.

Model 4: structured spatial effect, structured time effect, space-time interaction effects and a covariate.

Table 17: Results for various oesophageal models fitted with smoking as the covariate

Variables	Model 1	Model 2	Model 3	Model 4
Intercept (e^{β_0})	0.001	0.9578	0.0005	0.0005
Smoking (e^{β_1})	1.0121	1.0523	1.0121	1.0121
Year (e^{β_2})	-	0.0004	-	-
DIC	200.91	46067344	200.89	200.63

The multiplicative effect of smoking was observed to be $e^{\beta_1}=1.012$, indicating that oesophageal cancer was 1.2 % higher for smokers compared to non-smokers.

Table 17 presents the covariate estimates and DIC components for the four models: despite the added complexity due interaction between space and time, Model 4 was more plausible since it had the lowest DIC value. The relative risk values were obtained for the model as indicated in Table 18.

Table 18: The relative risks for counties with notified oesophageal cancer cases where smoking was the covariate

County	Relative Risks 2015/2016
Bomet	11.71
Embu	2.91
Kakamega	2.28
Kiambu	0.68
Machakos	0.99
Meru	6.68
Mombasa	1.09
Nairobi	1.78
Nakuru	2.59
Nyeri	4.01

Models where alcohol use was the covariate

In this section, four models were fitted similar to the four model above where alcohol use was the covariate.

Table 19: Results for various models fitted with alcohol use as the covariate

Variables	Model 1	Model 2	Model 3	Model 4
Intercept (e^{β_0})	0.0009	1.0725	0.0009	0.0009
Alcohol use (e^{β_1})	1.0346	1.0460	1.0346	1.0346
Year (e^{β_2})	-	0.0003	-	-
DIC	182.63	81715841	182.74	182.60

Table 19 presents the covariate estimates and DIC components for the four models: despite the added complexity due interaction between space and time, Model 4 was more plausible since it had the lowest DIC value. The multiplicative effect of alcohol use was observed to be $e^{\beta_1}=1.0346$, indicating that oesophageal cancer is 3.5 % higher to alcohol users as compared to non-alcohol users. Subsequently, relative risk values for the model were obtained as shown in Table 20 below.

Table 20: The relative risks for counties with notified oesophageal cancer cases with alcohol use as the covariate

County	Relative Risks 2015/2016
Bomet	11.75
Embu	2.80
Kakamega	2.43
Kiambu	0.64
Machakos	0.99
Meru	7.78
Mombasa	1.05
Nairobi	1.78
Nakuru	2.39
Nyeri	3.23

Spatial-temporal maps for oesophageal cancer model

Figure 4.6-4.7 shows the spatial-temporal distribution of the posterior estimates of the relative risks from 2015 to 2016 after accounting for spatially random and structured effects, temporal effects, space-time interactions and varying coefficient effects. These are interpreted as model-based relative risks. Counties with ($\mu_i > 1$) had higher than expected risk from a standard population while those with ($\mu_i < 1$) have lower than expected risk. In Figures 4.6-4.7 an increased risk can be seen in some parts of the country, characterized by a spatial relative risk above 1, and a posterior probabilities above 0.8.

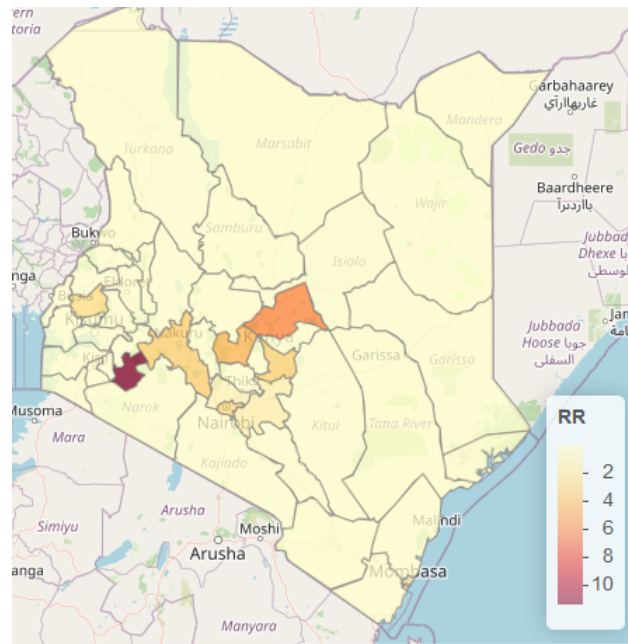


Figure 4.6: Spatial-temporal distribution of the relative risks for oesophageal cancer with smoking as the covariate.

Source: Author, 2021

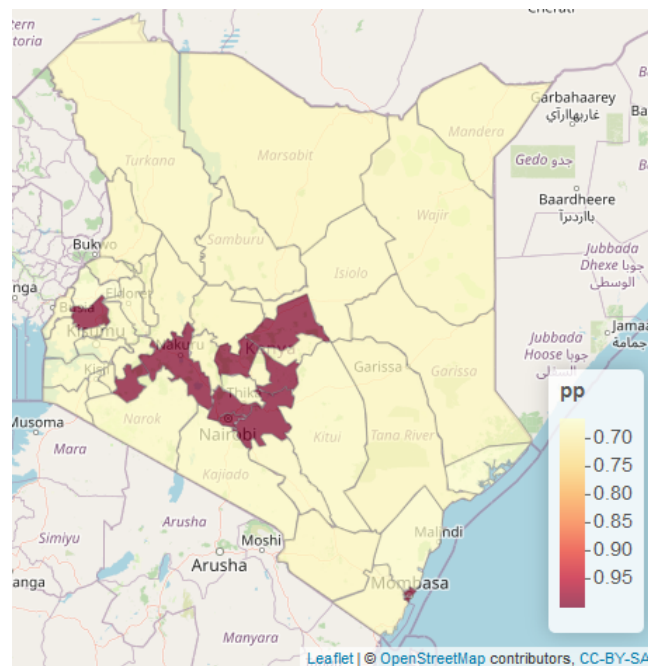


Figure 4.7: Map of the uncertainty for the spatial-temporal effects accounting for smoking effect (oesophageal cancer) $\mu_i : p(\mu_i > 1|y)$.

Source: Author, 2021

Oesophageal cancer models where alcohol use and smoking were covariates

Based on equation (3.31) three models incorporating various effects and covariates were fitted as follows:

Model 1: With structured, unstructured spatial effect, trend effects.

Model 2: With structured spatial effect, structured trend effect, global time effect as covariate.

Model 3: With structured spatial effect, structured time effect, space-time interaction, alcohol use, smoking, year and alcohol use-smoking interaction as the covariates.

Table 21: Results for various oesophageal models fitted with Alcohol, Smoking, Year and an interaction as the covariate

Variables	Model 1	Model 2	Model 3
Intercept (e^{β_0})	0.0012	0.9980	0.0036
Alcohol (e^{β_1})	1.0639	1.0555	1.1688
Smoking(e^{β_2})	0.0982	0.9910	1.0460
Year	-	0.0004	-
Alcohol*Smoking	-	-	0.9970
DIC	183.3621	73323521	173.3209

Table 21 presents the exponentiated covariate estimates and DIC components for the three models: the third model was the most plausible despite the added complexity due interaction between space and time as well as alcohol use-smoking interaction. The DIC value was 173.3209 which was the smallest. Since the interaction term coefficient value was less than 1, the conclusion was that, there was no interaction effect and hence proceeded to interpret the main effects. In this model the multiplicative effect of alcohol use was observed to be $e^{\beta_1}=1.1688$, indicating that oesophageal cancer was 16.88 % higher to alcohol users as compared

to non-alcohol users. The multiplicative effect of smoking was observed to be $e^{\beta_1}=1.0460$, indicating that oesophageal cancer was 4.60 % higher to smokers as compared to non-smokers.

Table 22: The relative risks for Model 3 where alcohol use and smoking were the covariates

County	Relative Risks
	2015/2016
Bomet	11.16
Embu	3.04
Kakamega	2.6
Kiambu	0.67
Machakos	1.05
Meru	5.93
Mombasa	1.09
Nairobi	2.18
Nakuru	2.88
Nyeri	3.89

The results in Table 22 revealed that Bomet, Meru, Nyeri, Embu, Nakuru, Kakamega and Nairobi counties had higher risk of oesophageal cancer respectively. The findings were consistent with model above where alcohol use and smoking variables were applied independently.

According to study findings in Figure 4.8, the light yellow shaded areas were low risk while the purple area exhibited the highest risk.

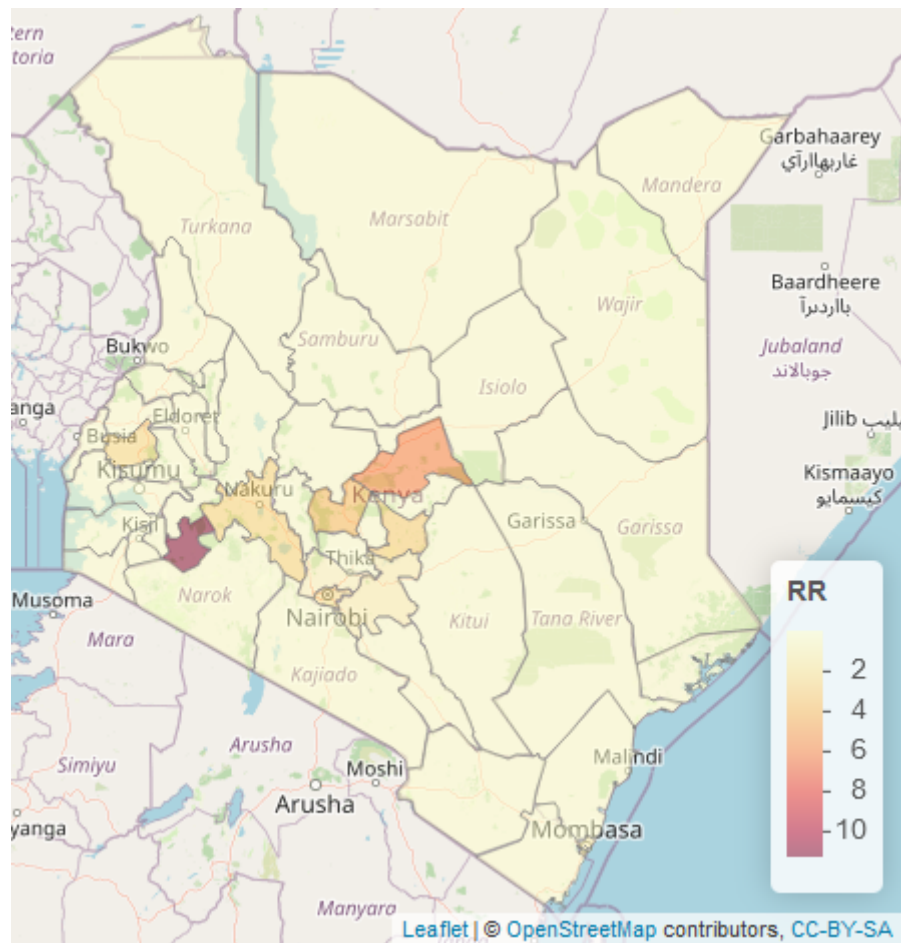


Figure 4.8: Spatial-temporal distribution of the relative risks for oesophageal cancer with alcohol use and smoking as the covariates.

Source: Author, 2021

4.4.4 Descriptive statistics for lung cancer

Table 23: Distribution of lung cancer by gender in 2015

Gender	Count of Gender	Percentage
Female	48	43.24
Male	63	56.74
Grand Total	111	100

According to data in Table 23, 63 (56.74%) of lung cancer cases were male while 48 (43.24%) of the cases were female.

Table 24: Distribution of lung cancer by gender in 2016

Gender	Count of Gender	Percentage
Female	63	43.15
Male	83	56.85
Grand Total	146	100

According to the data in Table 24, in 2016 83(56.85%) of lung cancer cases were male while 63(43.15%) of the cases were female.

4.4.5 Spatio-temporal models for lung cancer

Spatial-temporal model for lung cancer where smoking was the covariate

In this section, four models based on equation (3.31) were fitted, where smoking was the covariate. The models incorporated various spatial and temporal effects as follows:

Model 1: With structured, unstructured spatial effect, trend effects.

Model 2: With structured spatial effect, structured trend effect, global time effect and a covariate.

Model 3: With structured, unstructured spatial effects, structured trend effects and a covariate.

Model 4: structured spatial effect, structured time effect, space-time interaction effects and a covariate.

Table 25: Results for various models fitted with smoking as the covariate

Variables	Model 1	Model 2	Model 3	Model 4
Intercept (e^{β_0})	0.0327	0.5886	0.0327	0.0343
Smoking (e^{β_1})	1.3324	1.1996	1.3338	1.4021
Year (e^{β_2})	-	0.0612	-	-
DIC	129.55	211.78	129.47	127.12

Table 25 presents the covariate estimates and DIC components for the four models, Model 4 was selected since it had the lowest DIC value compared to others: The multiplicative effect of smoking was $e^{\beta_1}=1.4021$, indicating that lung cancer is 40.21 % higher to smokers as compared to non-smokers from the available data.

Table 26: The relative risks for counties with notified lung cancer cases with smoking as the covariate

County	Relative Risks 2015/2016
Bomet	0.68
Embu	5.01
Kakamega	0.19
Kiambu	1.99
Machakos	3.26
Meru	2.42
Mombasa	1.30
Nairobi	3.69
Nakuru	2.02
Nyeri	4.98

Relative risk greater than 1 indicated that the risk of developing lung cancer is higher in the specific counties than in the standard population. The relative risks in Table 16 indicated that majority of the counties where data was available had higher risk of developing lung cancer with exception of Bomet and Kakamega. In Figure 4.9 and 4.10 the light yellow coloured counties indicated low risk of lung cancer while the purple coloured counties indicated higher relative risk and probability value of more than 0.8.

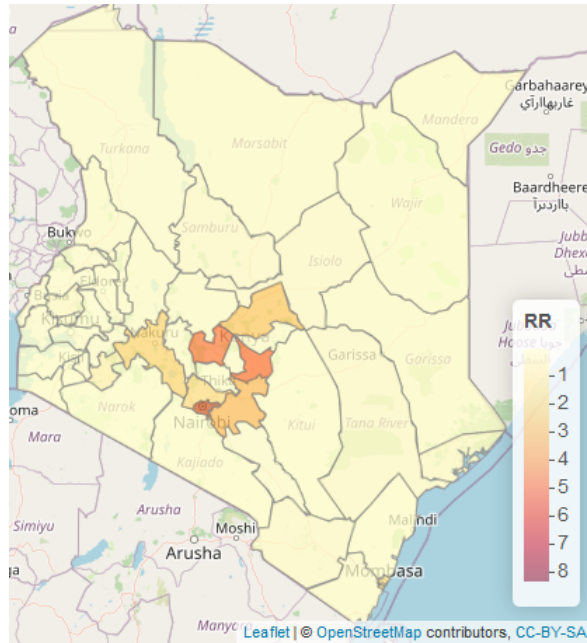


Figure 4.9: Spatial-temporal distribution of the relative risks for lung cancer with smoking as the covariate.

Source: Author, 2021

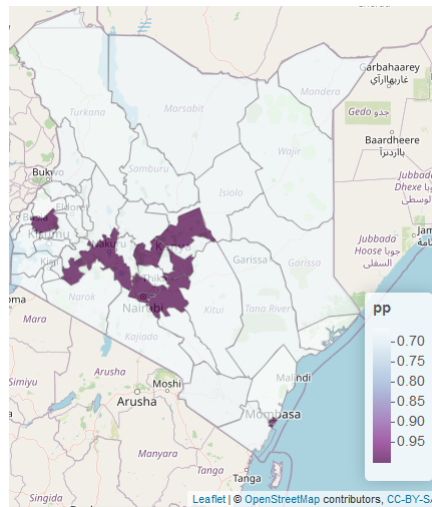


Figure 4.10: Map of the probability for the spatial temporal effects accounting for smoking effect (lung cancer) $\mu_i : p(\mu_i > 1|y)$.

Source: Author, 2021

Spatial-temporal model for lung cancer where alcohol use was the covariate

In this section, four models were fitted based on equation (3.31) where alcohol use was the covariate. The four models contained various effects as outlined below.

Model 1: With structured, unstructured spatial effect, trend effects.

Model 2: With structured spatial effect, structured trend effect, global time effect and a covariate.

Model 3: With structured, unstructured spatial effects, structured trend effects and a covariate.

Model 4: structured spatial effect, structured time effect, space-time interaction effects and a covariate.

Table 27: Results for various models fitted with alcohol use as the covariate

Variables	Model 1	Model 2	Model 3	Model 4
Intercept (e^{β_0})	0.0302	0.6344	0.0347	0.0342
Alcohol use (e^{β_1})	1.3689	0.05948	1.3716	1.3716
Year (e^{β_2})	-	1.1817	-	-
DIC	128.61	209.67	128.77	128.78

Table 27 presents the covariate estimates and DIC components for the four models, Model 1 was selected since it had the lowest DIC value compared to others: The study findings revealed, the multiplicative effect of alcohol use was $e^{\beta_1}=1.3689$, indicating that the risk of lung cancer is 36.89 % higher to alcohol users compared to non-alcohol users.

Table 28: The relative risks for counties with notified lung cancer cases where alcohol use is the covariate

County	Relative Risk 2015/2016
Bomet	0.69
Embu	5.00
Kakamega	0.19
Kiambu	1.78
Machakos	3.74
Meru	2.54
Mombasa	1.30
Nairobi	4.08
Nakuru	1.80
Nyeri	5.97

The relative risks in Table 28 indicated that in majority of the counties where the data was available the risk of developing lung cancer was higher than expected in the standard population since the relative risk values were greater than 1.

In Figure 4.11 the darker the colour the higher the relative risk as indicated in the RR bar. Nyeri, Embu, Nairobi and Machakos Counties had the highest risks respectively. The relative risk of the areas where the data was not available ranged between 0.0539 and 0.7971. In Figure 4.12 the purple shaded areas indicated a probability that the relative was greater than 1.

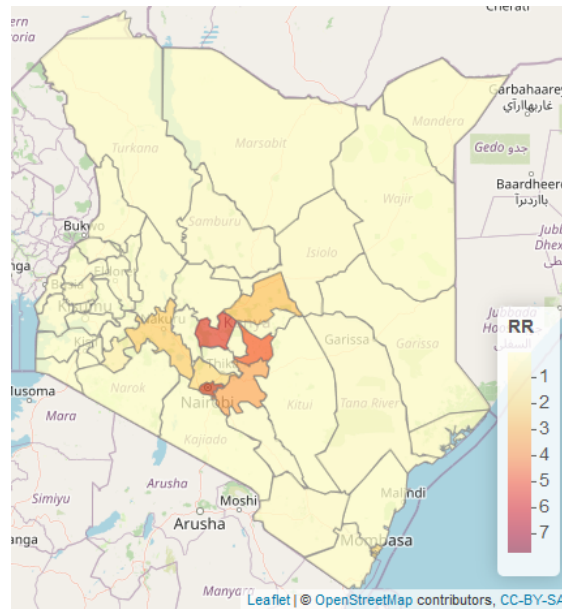


Figure 4.11: Spatial-temporal distribution of the relative risks for lung cancer with alcohol use as the covariate.

Source: Author, 2021

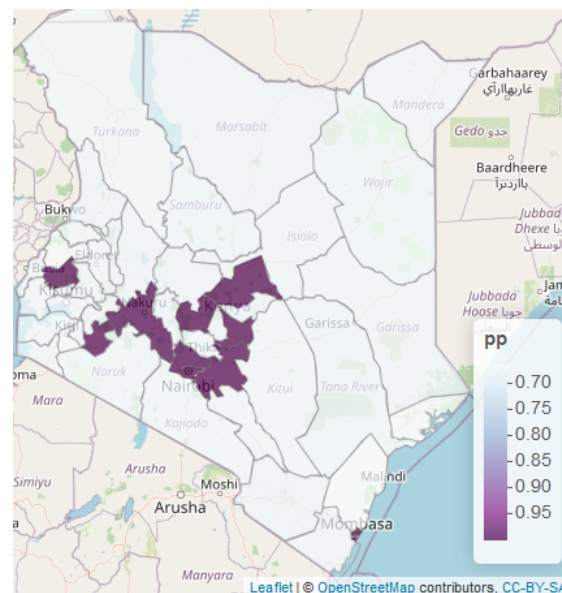


Figure 4.12: Map of the probability values accounting for alcohol use (lung cancer) $\mu_i : p(\mu_i > 1|y)$.

Source: Author, 2021

Lung cancer models where smoking and alcohol use were covariates

Three models based on equation (3.31) containing various effects as outlined below were applied.

Model 1: With structured, unstructured spatial effect, trend effects.

Model 2: With structured spatial effect, structured trend effect, global time effect and a covariate.

Model 3: structured spatial effect, structured time effect, space-time interaction effects, alcohol use, covariate, smoking and interaction covariate.

Table 29: Results for various lung cancer models fitted with alcohol use, smoking, year and an interaction as the covariate

Variables	Model 1	Model 2	Model 3
Intercept (e^{β_0})	0.0332	0.4700	0.0961
Alcohol (e^{β_1})	1.0919	0.6838	0.5477
Smoking(e^{β_2})	1.2361	1.8112	8.8640
Year	-	0.0755	-
Alcohol*Smoking(e^{β_3})	-	-	0.9277
DIC	131.014	203.088	127.059

Table 29 presents the covariate estimates and DIC components for the three models: the third model with a DIC value 127.059 was the most plausible despite the added complexity due interaction between space and time as well as smoking and alcohol use interaction term. The exponentiated coefficient value was 0.9277 which was less than 1 which means the interaction term was not significant. The findings in Model 3 revealed that the multiplicative effect of smoking was $e^{\beta_2}=8.8640$ indicating that smoking was a great risk factor for lung cancer in Kenya's counties. The multiplicative effect of alcohol use was observed to be $e^{\beta_1}=0.5477$ indicating that alcohol use was not a risk factor when modeled together with smoking. Relative risks for the model were obtained as shown in Table 20..

Table 30: The relative risks for counties with notified lung cancer cases where smoking and alcohol use were the covariates

County	Relative Risks 2015/2016
Bomet	0.44
Embu	0.9
Kakamega	0.10
Kiambu	1.85
Machakos	1.58
Meru	2.64
Mombasa	1.91
Nairobi	5.97
Nakuru	1.25
Nyeri	3.56

The study findings in Table 30 revealed that Nairobi, Nyeri, Meru, Kiambu and Mombasa had higher risk of lung cancer disease.

In Figure 4.13 the purple coloured area indicated the county with highest risk of lung cancer which was Nairobi while the light yellow coloured areas indicate the low risk areas.

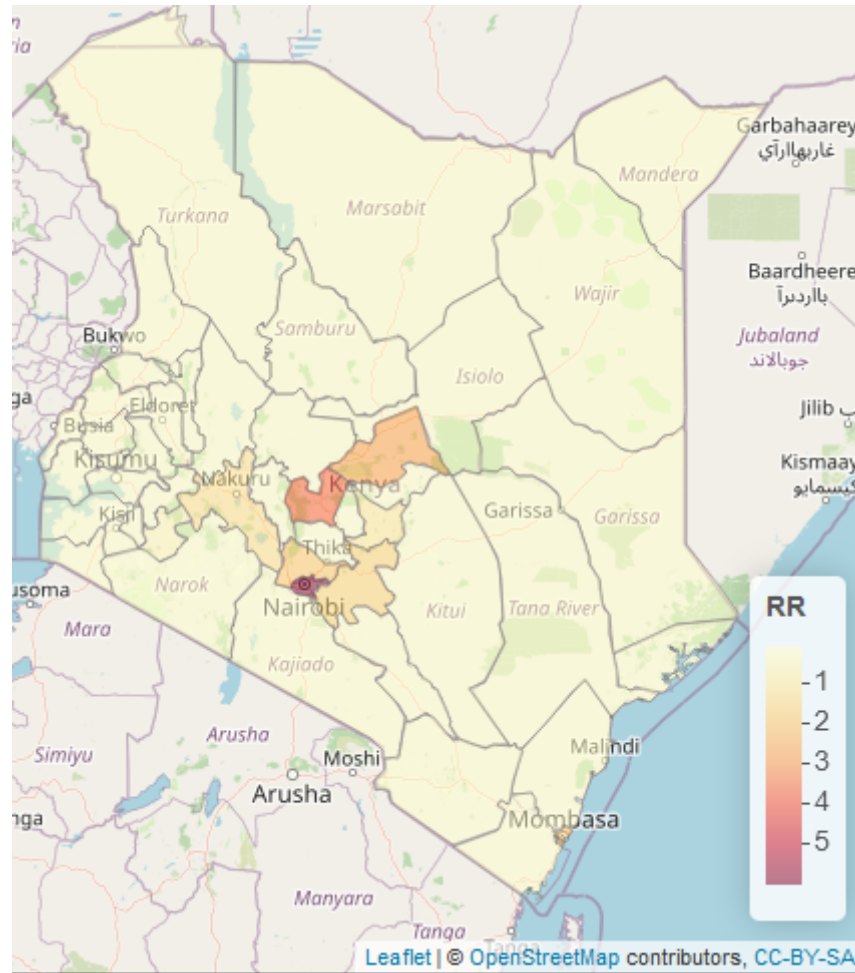


Figure 4.13: Spatial-temporal distribution of the relative risks for lung cancer with alcohol use-smoking as the covariate.

Source: Author, 2021

CHAPTER FIVE

5 SUMMARY, CONCLUSION AND RECOMMENDATIONS

5.1 Introduction

This chapter provides a summary of the major findings of this study and also sets to draw conclusions and make recommendations and suggestions for further research based on the results of this study.

5.2 Summary

Specific Objective 1: To model over-dispersion and conduct spatial correlations tests for cervical, oesophageal and lung cancer cases distribution in Kenya's counties.

Simple Poisson log normal regression models were not appropriate to model the three cancers due to over dispersion nature of the data sets. The spatial correlation tests revealed that there was no spatial auto correlation for the three types of cancer.

Specific Objective 2: To model cervical cancer cases using Poisson-Gamma and Spatial-temporal models.

The results revealed that counties where data was available among them Embu, Mombasa, Meru, Machakos and Nyeri counties had very high risk of cervical cancer. In counties where data was not available the model showed relative risks of cervical cancers was not very high but the risk was present, therefore spatial temporal models are very appropriate to estimate relative risks of diseases even when there is a small sample (and possibly an empty sample) in a given area by borrowing information from other neighboring regions.

Specific Objective 3: To model the effects of covariates on spatial-temporal distribution of oesophageal and lung cancer cases in Kenya's counties.

The study revealed that Bomet had highest relative risk of oesophageal cancer, followed by Meru, Nyeri, Embu, Nakuru, Kakamega Nairobi, Mombasa, Kiambu and Machakos counties respectively. Other counties had relatively low relative risks which ranged between 0.01-0.08, clearly even though the data was not available in these counties application of spatio-temporal accounting for covariates revealed that there was risk of oesophageal cancer in the counties. The study revealed that smoking and alcohol use are significant determinants of oesophageal cancer in Kenya. The study findings were consistent with Odera et al. (2017) who, identified alcohol drinking, genetic factors, dietary change/food preparation, and consumption of hot food as the main risk factors for oesophageal cancer. Patel et al. (2013) showed that there was positive and statistically significant relationship between tobacco smoking and development of oesophageal in Kenya, where in one study smokers had 2.51 odds of developing oesophageal cancer than non-smokers. Generation of spatio-temporal maps and identification of the risk factors from various counties with notified oesophageal cancer cases is a major milestone since previous studies on oesophageal cancer focused on specific regions. Previous studies had indicated that oesophageal cancer was more prevalent in western region of Kenya, but the study revealed that it is also prevalent in other counties such as Meru, Embu and Nyeri.

As per the study finding, it was evident that smoking and alcohol use were significant risk factors for lung cancer in Kenya. Meta-analyses by Bandera et al. (2001) indicated that the increased risk of lung cancer observed among alcoholics is mainly attributable to such residual confounding, since no consistent association was observed in never-smokers. Other risk factors include , indoor air pollution, which includes coal burning in poorly ventilated houses, burning of wood and other solid fuels, as well as fumes from high-temperature cooking using unrefined vegetable oils such as rapeseed oil, and occupational lung carcinogens such as

asbestos, silica, radon, heavy metals and poly-cyclic aromatic hydrocarbons. According to Malhotra et al. (2016), some of priorities for the prevention of lung cancer include control of occupational exposures, as well as indoor and outdoor air pollution, and understanding the carcinogenic and preventive effects of dietary and other lifestyle factors.

5.3 Recommendations

5.3.1 Recommendations for national and counties governments

We recommend that, since all counties had cervical cancer relative risk greater than 1, step up screening and avail vaccines to the appropriate groups.

The national, county and private health institutions should work closely to create awareness by disseminating information on oesophageal cancer and lung cancer especially in high risk areas as revealed by the study.

To mitigate oesophageal cancer, counties should create awareness on effects of smoking and alcohol use. In case of lung cancer, counties with relative risks greater than 1 should disseminate information elaborating the effects of smoking and alcohol use.

5.3.2 Recommendation for National Cancer Registry and other cancer registries

Despite success of this study, the biggest impediment in spatial temporal study was non-availability of adequate county data which could have provided more insight on the distribution of the the cancers in Kenya. Therefore the National Cancer Registry in collaboration with counties health departments should enhance cancer data collection to facilitate research and to inform the appropriate measures to be implemented to mitigate the increase of cancer cases.

5.3.3 Areas of further research

Further epidemiological studies can be conducted in areas with high relative risks to find out the other risk factors resulting to higher cases for the three cancers. The study obtain data for two years 2015 and 2016 from the National Cancer Registry, this was a short time series, further study can be conducted comprising longer periods of time to provide more insights and to compare the results with the ones obtained in this study. The study clearly revealed the relative risks for cervical, oesophageal and lung cancer in ten counties, subject to availability of more data the model can be applied in future studies to provide the relative risks for the remaining counties as well provide updated risks for the ten counties. Cancer is a generic term for various types of cancers, therefore, the study was not exhaustive of all cancers in Kenya, in the our continent and worldwide. Consequently, the models applied in this study can be replicated to model other diseases in the country and cancer in other regions.

References

- Adem, O. K., Mung'atu, K. H., Mwalili, S., Kibuchi, E., Ong'ang'o, J., and Sang, G. (2015). Spatial temporal modelling of tuberculosis in kenya using small area estimation.
- Agarwal, D. K., Gelfand, A. E., and Citron-Pousty, S. (2002). Zero-inflated models with application to spatial count data. *Environmental and Ecological statistics*, 9(4):341–355.
- Aloimonos, J. and Shulman, D. (1989). *Integration of Visual Modules*. Academic Press, New York.
- Amek, N., Bayoh, N., Hamel, M., Lindblade, K. A., Gimnig, J., Laserson, K. F., Slutsker, L., Smith, T., and Vounatsou, P. (2011). Spatio-temporal modeling of sparse geostatistical malaria sporozoite rate data using a zero inflated binomial model. *Spatial and spatio-temporal epidemiology*, 2(4):283–290.
- Anderson, C., Lee, D., and Dean, N. (2014). Identifying clusters in bayesian disease mapping. *Biostatistics*, 15(3):457–469.
- Anderson, C. and Ryan, L. (2017). A comparison of spatio-temporal disease mapping approaches including an application to ischaemic heart disease in new south wales, australia. *International journal of environmental research and public health*, 14(2):146.
- Bandera, E. V., Freudenheim, J. L., and Vena, J. E. (2001). Alcohol consumption and lung cancer: a review of the epidemiologic evidence. *Cancer Epidemiology and Prevention Biomarkers*, 10(8):813–821.
- Benavent, R. and Morales, D. (2016). Multivariate fay–herriot models for small area estimation. *Computational Statistics & Data Analysis*, 94:372–390.
- Bergen, J. R. and Landy, M. S. (1991). Computational modeling of visual texture segregation. In Landy, M. S. and Movshon, J. A., editors, *Computational Models of Visual Processing*, pages 253–271. MIT Press, Cambridge, MA.
- Bernardinelli, L., Clayton, D., Pascutto, C., Montomoli, C., Ghislandi, M., and Songini, M. (1995). Bayesian analysis of space-time variation in disease risk. *Statistics in medicine*, 14(21-22):2433–2443.
- Besag, J., York, J., and Mollié, A. (1991). Bayesian image restoration, with two applications in spatial statistics. *Annals of the institute of statistical mathematics*, 43(1):1–20.
- Bivand, R. (2019). Creating neighbours.
- Bivand, R., Gómez-Rubio, V., and Rue, H. (2015). Spatial data analysis with r-inla with some extensions. American Statistical Association.
- Blangiardo, M. and Cameletti, M. (2015). *Spatial and spatio-temporal Bayesian models with R-INLA*. John Wiley & Sons.

- Braunstein, M. L. (1968). Motion and texture as sources of slant information. *Journal of Experimental Psychology*, 78:247–253.
- Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R. L., Torre, L. A., and Jemal, A. (2018). Global cancer statistics 2018: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: a cancer journal for clinicians*, 68(6):394–424.
- Brezger, A., Kneib, T., and Lang, S. (2003). Bayesx: Analysing bayesian structured additive regression models. Technical report, Discussion paper//Sonderforschungsbereich 386 der Ludwig-Maximilians àS.
- Cancer.Net (2021). Esophageal cancer symptoms and signs. <https://www.cancer.net/cancer-types/esophageal-cancer/symptoms-and-signs>.
- Carvalho, J. R. P. D., Nakai, A. M., and Monteiro, J. E. (2016). Spatio-temporal modeling of data imputation for daily rainfall series in homogeneous zones. *Revista Brasileira de Meteorologia*, 31(2):196–201.
- Chandra, H. (2003). Overview of small area estimation techniques. *Indian Agricultural Statistics Research Institute*, pages 75–88.
- Charras-Garrido, M., Abrial, D., Goër, J. D., Dachian, S., and Peyrard, N. (2012). Classification method for disease risk mapping based on discrete hidden markov random fields. *Biostatistics*, 13(2):241–255.
- Cliff, A. D. and Ord, K. (1970). Spatial autocorrelation: a review of existing and new measures with applications. *Economic Geography*, 46(sup1):269–292.
- Cnaan, A., Laird, N. M., and Slasor, P. (1997). Using the general linear mixed model to analyse unbalanced repeated measures and longitudinal data. *Statistics in medicine*, 16(20):2349–2380.
- Davis, H. T., Aelion, C. M., McDermott, S., and Lawson, A. B. (2009). Identifying natural and anthropogenic sources of metals in urban and rural soils using gis-based data, pca, and spatial interpolation. *Environmental Pollution*, 157(8-9):2378–2385.
- Diggle, P. and Lophaven, S. (2006). Bayesian geostatistical design. *Scandinavian Journal of Statistics*, 33(1):53–64.
- Dosher, B. A., Landy, M. S., and Sperling, G. (1989a). The kinetic depth effect and optic flow — I. 3D shape from fourier motion. *Vision Research*, 29:1789–1813.
- Dosher, B. A., Landy, M. S., and Sperling, G. (1989b). Ratings of kinetic depth in multi-dot displays. *Journal of Experimental Psychology: Human Perception and Performance*, 15:816–825.
- Dosher, B. A., Sperling, G., and Landy, M. S. (1989c). The kinetic depth — A nonexistent article. *Vision Research*, 29:100–113.
- Dosher, B. A., Sperling, G., and Landy, M. S. (1990). The kinetic depth — Another nonexistent article. *Vision Research*, 29:100–113.

- Epstein, M., Achong, B., and Barr, Y. (1964). Virus particles in cultured lymphoblasts from burkitt's lymphoma. *The Lancet*, 283(7335):702–703.
- Esin, A. (2018). Flexibility of using com-poisson regression model for count data. *Statistics, Optimization & Information Computing*, 6(2):278–285.
- Fay III, R. E. and Herriot, R. A. (1979). Estimates of income for small places: an application of james-stein procedures to census data. *Journal of the American Statistical Association*, 74(366a):269–277.
- Ferlay, J., Soerjomataram, I., Dikshit, R., Eser, S., Mathers, C., Rebelo, M., Parkin, D. M., Forman, D., and Bray, F. (2015). Cancer incidence and mortality worldwide: sources, methods and major patterns in globocan 2012. *International journal of cancer*, 136(5):E359–E386.
- Fontham, E. T., Wolf, A. M., Church, T. R., Etzioni, R., Flowers, C. R., Herzig, A., Guerra, C. E., Oeffinger, K. C., Shih, Y.-C. T., Walter, L. C., et al. (2020). Cervical cancer screening for individuals at average risk: 2020 guideline update from the american cancer society. *CA: A Cancer Journal for Clinicians*, 70(5):321–346.
- for Research on Cancer, I. A. et al. Iarc monographs, chemical agents and related occupations. vol 100f (2012)[displayed 13 october 2017].
- Foreman, K. J., Lozano, R., Lopez, A. D., and Murray, C. J. (2012). Modeling causes of death: an integrated approach using codem. *Population health metrics*, 10(1):1–23.
- Freudenheim, J. L., Ritz, J., Smith-Warner, S. A., Albanes, D., Bandera, E. V., Van Den Brandt, P. A., Colditz, G., Feskanich, D., Goldbohm, R. A., Harnack, L., et al. (2005). Alcohol consumption and risk of lung cancer: a pooled analysis of cohort studies-. *The American journal of clinical nutrition*, 82(3):657–667.
- Frying, H.-t. Volume 95 household use of solid fuels and high-temperature frying.
- Gelfand, A. E., Diggle, P., Guttorp, P., and Fuentes, M. (2010). *Handbook of spatial statistics*. CRC press.
- Gómez-Rubio, V., Best, N., Richardson, S., Li, G., and Clarke, P. (2010). Bayesian statistics small area estimation.
- Gonzalez, M. E. and Waksberg, J. (1973). *Estimation of the error of synthetic estimates*. US Census Bureau [custodian].
- Horner, M.-J., Altekruse, S. F., Zou, Z., Wideroff, L., Katki, H. A., and Stinchcomb, D. G. (2011). Us geographic distribution of prevaccine era cervical cancer screening, incidence, stage, and mortality. *Cancer Epidemiology and Prevention Biomarkers*, 20(4):591–599.
- Jiang, J. and Lahiri, P. (2006). Mixed model prediction and small area estimation. *Test*, 15(1):1.

- Kasper, D., Fauci, A., Hauser, S., Longo, D., Jameson, J., and Loscalzo, J. (2015). *Harrison's principles of internal medicine, 19e*, volume 1. Mcgraw-hill.
- Kass, R. E. and Raftery, A. E. (1995). Bayes factors. *Journal of the american statistical association*, 90(430):773–795.
- Khana, D., Rossen, L. M., Hedegaard, H., and Warner, M. (2018a). A bayesian spatial and temporal modeling approach to mapping geographic variation in mortality rates for subnational areas with r-inla. *Journal of data science: JDS*, 16(1):147.
- Khana, D., Rossen, L. M., Hedegaard, H., and Warner, M. (2018b). A bayesian spatial and temporal modeling approach to mapping geographic variation in mortality rates for subnational areas with r-inla. *Journal of data science: JDS*, 16(1):147.
- Knorr-Held, L. and Rasser, G. (1999). Bayesian detection of clusters and discontinuities in disease maps: Simulations.(revised, june 1999).
- Korir, A., Okerosi, N., Ronoh, V., Mutuma, G., and Parkin, M. (2015). Incidence of cancer in n airobi, k enya (2004–2008). *International journal of cancer*, 137(9):2053–2059.
- Korte, J. E., Brennan, P., Henley, S. J., and Boffetta, P. (2002). Dose-specific meta-analysis and sensitivity analysis of the relation between alcohol consumption and lung cancer risk. *American journal of epidemiology*, 155(6):496–506.
- Lawson, A. B. (2013a). *Bayesian disease mapping: hierarchical modeling in spatial epidemiology*. Chapman and Hall/CRC.
- Lawson, A. B. (2013b). *Bayesian disease mapping: hierarchical modeling in spatial epidemiology*. Chapman and Hall/CRC.
- Lawson, A. B., Biggeri, A., Boehning, D., Lesaffre, E., Viel, J.-F., Clark, A., Schlattmann, P., and Divino, F. (2000). Disease mapping models: an empirical evaluation. disease mapping collaborative group. *Statistics in medicine*, 19(17):2217–41.
- Lawson, A. B. and Rotejanaprasert, C. (2014). Childhood brain cancer in florida: a bayesian clustering approach. *Statistics and Public Policy*, 1(1):99–107.
- Lee, D. (2013). Carbayes: an r package for bayesian spatial modeling with conditional autoregressive priors. *Journal of Statistical Software*, 55(13):1–24.
- Lee, D., Ferguson, C., and Mitchell, R. (2009). Air pollution and health in scotland: a multicity study. *Biostatistics*, 10(3):409–423.
- Lee, D., Minton, J., and Pryce, G. (2015). Bayesian inference for the dissimilarity index in the presence of spatial autocorrelation. *Spatial Statistics*, 11:81–95.
- Lee, D., Rushworth, A., and Napier, G. (2018). Spatio-temporal areal unit modelling in r with conditional autoregressive priors using the carbayesst package. *Journal of Statistical Software*, 84(9).

- Leroux, B. G., Lei, X., and Breslow, N. (2000). Estimation of disease rates in small areas: a new mixed model for spatial dependence. In *Statistical models in epidemiology, the environment, and clinical trials*, pages 179–191. Springer.
- Lindgren, F., Rue, H., et al. (2015). Bayesian spatial modelling with r-inla. *Journal of Statistical Software*, 63(19):1–25.
- Malhotra, J., Malvezzi, M., Negri, E., La Vecchia, C., and Boffetta, P. (2016). Risk factors for lung cancer worldwide. *European Respiratory Journal*, 48(3):889–902.
- Martino, S. and Rue, H. (2009). Implementing approximate bayesian inference using integrated nested laplace approximation: A manual for the inla program. *Department of Mathematical Sciences, NTNU, Norway*.
- Martins, T. G., Simpson, D., Lindgren, F., and Rue, H. (2013a). Bayesian computing with inla: new features. *Computational Statistics & Data Analysis*, 67:68–83.
- Martins, T. G., Simpson, D., Lindgren, F., and Rue, H. (2013b). Bayesian computing with inla: new features. *Computational Statistics & Data Analysis*, 67:68–83.
- Mathers, C. D., Ma Fat, D., Inoue, M., Rao, C., and Lopez, A. D. (2005). Counting the dead and what they died from: an assessment of the global status of cause of death data. *Bulletin of the world health organization*, 83:171–177c.
- Miaou, S.-P. and Lord, D. (2003). Modeling traffic crash-flow relationships for intersections: dispersion parameter, functional form, and bayes versus empirical bayes methods. *Transportation Research Record*, 1840(1):31–40.
- Moraga, P. (2018). Small area disease risk estimation and visualization using r. *R J*, 10:495–506.
- Moran, P. A. (1950). A test for the serial independence of residuals. *Biometrika*, 37(1/2):178–181.
- Mwangi, K. J. (2014). *Use of GIS in mapping of cancer prevalence a case study of Uasin Gishu County*. PhD thesis, UNIVERSITY OF NAIROBI.
- Myer, M. H., Campbell, S. R., and Johnston, J. M. (2017). Spatiotemporal modeling of ecological and sociological predictors of west nile virus in suffolk county, ny, mosquitoes. *Ecosphere*, 8(6):e01854.
- Nazia, N., Ali, M., Jakariya, M., Nahar, Q., Yunus, M., and Emch, M. (2018). Spatial and population drivers of persistent cholera transmission in rural bangladesh: Implications for vaccine and intervention targeting. *Spatial and spatio-temporal epidemiology*, 24:1–9.
- Neelon, B., Ghosh, P., and Loebs, P. F. (2013). A spatial poisson hurdle model for exploring geographic variation in emergency department visits. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 176(2):389–413.

- Neyens, T., Faes, C., and Molenberghs, G. (2012). A generalized poisson-gamma model for spatially overdispersed data. *Spatial and spatio-temporal epidemiology*, 3(3):185–194.
- Odera, J. O., Odera, E., Githangâa, J., Walong, E. O., Li, F., Xiong, Z., and Chen, X. L. (2017). Esophageal cancer in kenya. *American journal of digestive disease*, 4(3):23.
- Oleson, J. J. and Wikle, C. K. (2013). Predicting infectious disease outbreak risk via migratory waterfowl vectors. *Journal of Applied Statistics*, 40(3):656–673.
- Pacella-Norman, R., Urban, M., Sitas, F., Carrara, H., Sur, R., Hale, M., Ruff, P., Patel, M., Newton, R., Bull, D., et al. (2002). Risk factors for oesophageal, lung, oral and laryngeal cancers in black south africans. *British journal of cancer*, 86(11):1751–1756.
- Parkin, D. M., Bray, F., Ferlay, J., and Jemal, A. (2014). Cancer in africa 2012. *Cancer epidemiology, biomarkers & prevention : a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology*, 23:953–66.
- Patel, K., Wakhisi, J., Mining, S., Mwangi, A., and Patel, R. (2013). Esophageal cancer, the topmost cancer at mtrh in the rift valley, kenya, and its potential risk factors. *International Scholarly Research Notices*, 2013.
- Pirani, M., Gulliver, J., Fuller, G. W., and Blangiardo, M. (2014). Bayesian spatiotemporal modelling for the assessment of short-term exposure to particle pollution in urban areas. *Journal of Exposure Science and Environmental Epidemiology*, 24(3):319.
- Pringle, D. (1995). Disease mapping: A comparative analysis of maximum likelihood and empirical bayes estimates of disease risk. *The Economic and Social Review*.
- Rachmawati, R. N., Djuraidah, A., Fitrianto, A., and Sumertajaya, I. M. (2018). Spatio-temporal models using r-inla with generalized extreme value distribution in hierarchical bayes regression.
- Rao, J. (2003). Some new developments in small area estimation. *Journal of the Iranian Statistical Society*, 2(2):145–169.
- Rue, H. and Martino, S. (2007). Approximate bayesian inference for hierarchical gaussian markov random field models. *Journal of statistical planning and inference*, 137(10):3177–3192.
- Rue, H., Martino, S., and Chopin, N. (2009). Approximate bayesian inference for latent gaussian models by using integrated nested laplace approximations. *Journal of the royal statistical society: Series b (statistical methodology)*, 71(2):319–392.
- Saei, A. and Chambers, R. (2005). Empirical best linear unbiased prediction for out of sample areas.

- Schaafsma, T., Wakefield, J., Hanisch, R., Bray, F., Schüz, J., Joy, E. J., Watts, M. J., and McCormack, V. (2015). Africaâs oesophageal cancer corridor: geographic variations in incidence correlate with certain micronutrient deficiencies. *PloS one*, 10(10):e0140107.
- Schaible, W. L. (1996). Recommendations and cautions. In *Indirect Estimators in US Federal Programs*, pages 188–195. Springer.
- Schottenfeld, D. and Fraumeni Jr, J. F. (2006). *Cancer epidemiology and prevention*. Oxford University Press.
- Schrödle, B. and Held, L. (2011). Spatio-temporal disease mapping using inla. *Environmetrics*, 22(6):725–734.
- Society, A. C. (2021). Lung cancer risk factors. <https://www.cancer.org/cancer/lung-cancer/causes-risks-prevention/risk-factors.html>.
- Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the royal statistical society: Series b (statistical methodology)*, 64(4):583–639.
- StataCorp, L. (2013). Stata multilevel mixed-effects reference manual. *College Station, TX: StataCorp LP*.
- Stern, H. and Cressie, N. A. (1999). Inference for extremes in disease mapping.
- Tenge, C., Kuremu, R., Buziba, N., Patel, K., and Were, P. (2009). Burden and pattern of cancer in western kenya. *East African medical journal*, 86(1).
- Vidoni, C., Vallino, L., Ferraresi, A., Secomandi, E., Salwa, A., Chinthakindi, M., Galetto, A., Dhanasekaran, D. N., and Isidoro, C. (2021). Epigenetic control of autophagy in womenâs tumors: role of non-coding rnas. *Journal of Cancer Metastasis and Treatment*, 7.
- Wakefield, J. (2007). Disease mapping and spatial regression with count data. *Biostatistics*, 8(2):158–183.
- WHO (2020). Human papillomavirus (hpv) and cervical cancer. [https://www.who.int/en/news-room/fact-sheets/detail/human-papillomavirus-\(hpv\)-and-cervical-cancer](https://www.who.int/en/news-room/fact-sheets/detail/human-papillomavirus-(hpv)-and-cervical-cancer).
- Wikle, C. K. and Anderson, C. J. (2003). Climatological analysis of tornado report counts using a hierarchical bayesian spatiotemporal model. *Journal of Geophysical Research: Atmospheres*, 108(D24).
- Yanosky, J. D., Paciorek, C. J., Laden, F., Hart, J. E., Puett, R. C., Liao, D., and Suh, H. H. (2014). Spatio-temporal modeling of particulate air pollution in the conterminous united states using geographic and meteorological predictors. *Environmental Health*, 13(1):63.
- Yue, Y. R. and Wang, X.-F. (2014). Spatial gaussian markov random fields: Modelling, applications and efficient computations.

A APPENDICES

APPENDIX A: R-Codes for various objectives.

Assessing the over dispersion and spatial correlation of cervical cancer cases

```

library(MASS)
library(sf)
library(maptools)
library(spdep)
library(maptools)
library(raster)
library(plyr) library(ggplot2)
library(rgdal)
library(SpatialEpi)
library("CARBayesdata")
library("sp")
Kenya<-getData("GADM", country="KE", level=0)
Kenya3<-getData("GADM", country="KE", level=1)
Kenya3
NAME_1<-Kenya3$NAME_1
NAME_1
CancerK4<-read.csv("D:\\JOSEPH KURIA FOLDER\\PHD 2019 TEX\\
November 2019\\
CORRECT DURING PREPARATION DOCUMENTS September 2021\\
Appendix R-Code\\Cervixdata.csv")
CancerK4

```

```

Y7<-CancerK4$Cases Y7
Pop1<-CancerK4$Pop
Pop1 E7<-expected(Pop1,Y7, 1)
CancerK4$SIR<-Y7/E7
CancerK4$SIR
#
library("dplyr")
SIR.av2 <- summarise(group_by(CancerK4,NAME_1), SIR.mean2

=mean(SIR))
SIR.av2
SIR.av2<-as.data.frame(SIR.av2)
SIR.av2
NAME_1<-as.character(SIR.av2$NAME_1)
NAME_1
SIR.mean2<-SIR.av2$SIR.mean2
SIR.mean2
SIR.av2<-data.frame(NAME_1,SIR.mean2)
SIR.av2
ID<-as.character(seq(1,47))
library("spdep")
W.nb <- poly2nb(Kenya3, row.names = SIR.av2$NAME_1)
W.list <- nb2listw(W.nb, style = "B")
W <- nb2mat(W.nb, style = "B")
formula<- Y7 ~ offset(log(E7))
model.c <- glm(formula = formula, family = "poisson",data=CancerK4)
summary(model.c)
library(AER)
dispersiontest (model.c)

```



```

#Computing Moran I statistic
resid.glm <- residuals(model.c)
resid.glm
summary(model.c)$coefficients
moran.mc(x = resid.glm[1:47], listw = W.list, nsim = 10000)

```

Assessing the over dispersion and spatial correlation of oesophageal cancer cases

```

library(MASS)
library(sf)
library(maptools)
library(spdep)
library(maptools)
library(raster)
library(plyr) library(ggplot2)
library(rgdal)
library(SpatialEpi)
library("CARBayesdata")
library("sp")
Kenya<-getData("GADM", country="KE", level=0)
Kenya3<-getData("GADM", country="KE", level=1)
Kenya3
NAME_1<-Kenya3$NAME_1 NAME_1
OECdata<-read.csv("D:\\JOSEPH KURIA FOLDER\\PHD 2019 TEX\\
DatasetsandRcodes\\OECANCER 2021.csv")
OECdata
Y5<-OECdata$Cases Y5
Pop1<-OECdata$Pop
Pop1 E15<-expected(Pop1,Y5, 1)

```

```

OECdata$SIR<-Y5/E15
OECdata$SIR
#####
library("dplyr")
SIR.av2 <- summarise(group_by(OECdata,NAME_1), SIR.mean2 =mean(SIR))
SIR.av2
SIR.av2<-as.data.frame(SIR.av2)
SIR.av2
NAME_1<-as.character(SIR.av2$NAME_1)
NAME_1
SIR.mean2<-SIR.av2$SIR.mean2
SIR.mean2
SIR.av2<-data.frame(NAME_1,SIR.mean2)
SIR.av2
ID<-as.character(seq(1,47))
library("spdep")
W.nb <- poly2nb(Kenya3, row.names = SIR.av2$NAME_1)
W.list <- nb2listw(W.nb, style = "B")
W <- nb2mat(W.nb, style = "B")
formula13 <- Y5 ~ offset(log(E15))
model3 <- glm(formula = formula13, family = "poisson")
summary(model3)
library(AER)
dispersiontest(model3)
#Computing Moran I statistic
resid.glm <- residuals(model3)
resid.glm
summary(model3)$coefficients
moran.mc(x = resid.glm[1:47], listw = W.list, nsim = 10000)

```

Assessing the over dispersion and spatial correlation of lung cancer data

```
library(MASS)
library(sf)
library(maptools)
library(spdep)
library(maptools)
library(raster)
library(plyr)
library(ggplot2)
library(rgdal)
library(SpatialEpi)
library("CARBayesdata")
library("sp")
Kenya<-getData("GADM", country="KE", level=0)
Kenya3<-getData("GADM", country="KE", level=1)
Kenya3
NAME_1<-Kenya3$NAME_1
NAME_1
LCdata<-read.csv("D:\\JOSEPH KURIA FOLDER\\PHD 2019 TEX\\
DatasetsandRcodes\\Lungcancer Data.csv")
LCdata
Y6<-LCdata$Cases
Y6
Pop3<-LCdata$Pop
Pop3
L16<-expected(Pop3,Y6, 1)
LCdata$SIR<-Y6/L16
LCdata$SIR
####
```

```

library("dplyr")
SIR.av3 <- summarise(group_by(LCdata,NAME_1), SIR.mean3 =

mean(SIR))
SIR.av3
SIR.av3<-as.data.frame(SIR.av3)
SIR.av3
NAME_1<-as.character(SIR.av3$NAME_1)
NAME_1
SIR.mean3<-SIR.av3$SIR.mean3
SIR.mean3
SIR.av3<-data.frame(NAME_1,SIR.mean3)
SIR.av3
library("spdep")
W.nb <- poly2nb(Kenya3, row.names = SIR.av3$NAME_1)
W.list <- nb2listw(W.nb, style = "B")
W <- nb2mat(W.nb, style = "B")
formula6 <- Y6 ~ offset(log(L16))
model6 <- glm(formula = formula6, family = "poisson")
summary(model6)
library(MASS)
library(AER)
dispersiontest(model6)
#####
resid.glm <- residuals(model6)
resid.glm summary(model6)$coefficients
morán.mc(x = resid.glm[1:47], listw = W.list, nsim = 10000)

```

To model cervical cancer cases using Poisson-Gamma and Spatial-temporal models.

Poisson-gamma model for cervical cancer cases

```

library(MASS)
library(sf)
library(maptools)
library(spdep)
library(maptools)
library(raster)
library(plyr)
library(ggplot2)
library(rgdal)
library(SpatialEpi)
library("CARBayesdata")
library("sp")
Kenya<-getData("GADM", country="KE", level=0)
Kenya3<-getData("GADM", country="KE", level=1)
Kenya3
NAME_1<-Kenya3$NAME_1
NAME_1
#Reading in data
CancerK4<-read.csv("D:\\JOSEPH KURIA FOLDER\\PHD 2019 TEX\\
November 2019
\\CORRECT DURING PREPARATION DOCUMENTS September 2021\\
Appendix R-Code\\Cervixdata.csv")
head(CancerK4)
CancerK4 Cases<-CancerK4$Cases
Cases

```

```

Pop<-CancerK4$Pop
#calculating expected cancer cases per county
E<-expected(Pop,Cases, 1)
#generating Standardized Incidence Rates
CancerK4$SIR<-Cases/E
CancerK4$SIR
Cervix<-data.frame(NAME_1,Cases,E)
Cervix$SIR<-Cases/E
Cervix$SIR Cervix
#Histogram of cervical cancer cases
hist(Cases, col="grey", border=NA, las=1,xlab="Cervical cancer cases",

main="Histogram of cervical cancer cases")
#Generating Poisson-Gamma model
library("hglm")
pois.gamma <-hglm(fixed = Cases~offset(log(E)), random = ~1 |

NAME_1,fix.disp = 1, family =poisson(), rand.family =

Gamma(link = log), data = Cervix,calc.like = TRUE)

summary(pois.gamma)
print(pois.gamma,print.ranef = TRUE)

```

Spatial-temporal models for cervical cancer cases

```

library(MASS)
library(sf)
library(maptools)
library(spdep)

```

```

library(raster)
library(plyr)
library(ggplot2)
library(rgdal)
library(SpatialEpi)
library("sp")
Kenya<-getData("GADM", country="KE", level=0)
Kenya3<-getData("GADM", country="KE", level=1)
Kenya3
NAME_1<-Kenya3$NAME_1
SCdata<-read.csv("D:\\JOSEPH KURIA FOLDER\\PHD 2019 TEX\\
November 2019
\\STcervixdata.csv")
head(SCdata)
Exp2015C<-expected(SCdata$Pop,SCdata$Obs2015, 1)
Exp2015C
Exp2016C<-expected(SCdata$Pop,SCdata$Obs2016, 1)
Exp2016C
SCdata$Exp2015C<-expected(SCdata$Pop,SCdata$Obs2015, 1)
SCdata$Exp2015C
SCdata$Exp2016C<-expected(SCdata$Pop,SCdata$Obs2016, 1)
SCdata$Exp2016C SCdata
Kenya3 <- merge(Kenya3, SCdata)
Kenya3
low.vector <- as.vector(as.matrix(SCdata[,2:3]))#by column
low.vector
E.vector <- as.vector(as.matrix(SCdata[,5:6]))#by column
E.vector
year <- numeric(0)

```

```

for(i in 1:2){
year<- append(year,rep(i,dim(SCdata)[1]))
}
year
NAME_1<- as.factor(rep(SCdata[,1],2))
NAME_1
SCdata
dataSC<- data.frame(NAME_1,y= low.vector,E= E.vector,
ID.area=as.numeric(NAME_1), ID.area1=as.numeric(NAME_1),
year=year, ID.year = year, ID.year1=year,
ID.area.year = seq(1,length(NAME_1)))
dataSC
####SIR
datanew<-data.frame(NAME_1,y= low.vector,E= E.vector)
datanew
SIR<-datanew$y/datanew$E SIR
library("dplyr")
datanew<-data.frame(NAME_1,y= low.vector,E= E.vector,SIR)
datanew SIR.av <- summarise(group_by(datanew,NAME_1 ),
SIR.mean =mean(SIR))
SIR.av
SIR.av<-as.data.frame(SIR.av)
SIR.av
####
SIR.av1<-read.csv("D:\\JOSEPH KURIA FOLDER\\
PHD 2019 TEX\\DatasetsandRcodes
\\SIRCERVICAL.csv")
SIR.av1
SIR<-SIR.av1$SMR.mean

```



```

Kenya3@data
Kenya3@data$SIR <- SIR
Kenya3@data$SIR
#Standardized Incidence Rates
library(leaflet)
pal <- colorNumeric(palette = "YlOrRd", domain = SIR)
pal
labels <- sprintf("<strong> %s </strong> <br/>")
Pop: %s <br/> SIR: %s ",
Kenya3$NAME_1,
Pop, round(SIR, 2)
)%>% lapply(htmltools::HTML)
lSIR<- leaflet(Kenya3) %>%
addTiles() %>%
addPolygons( color = "grey", weight = 1, fillColor = ~
pal(SIR),
fillOpacity = 0.7, highlightOptions = highlightOptions
(weight = 4),
label = labels,
labelOptions = labelOptions(
noHide = FALSE,
style =
list(
"font-weight" = "normal",
padding = "3px 8px" ),
textsize = "15px", direction = "auto"
)
) %>%
addLegend(

```

```

pal = pal, values = ~Kenya3@data$SIR , opacity = 0.7,
title = "SIR",
position = "bottomright" )
lSIR

###Models
library(spdep)
library(INLA)
nb <- poly2nb(Kenya3)
nb2INLA("Kenya3.adj", nb)
g <- inla.read.graph(filename = "Kenya3.adj")
# Models without covariates
#Model 1 with Space (Structured), Unstructured, Time
formula1 <- y ~ 1+ f(ID.area,model="bym",graph=g)+
f(ID.area1,year,model="iid")+year

Model1 <- inla(formula1,family="poisson",data=dataSC
,E=E,control.predictor=list(compute=TRUE),
control.compute=list(dic=TRUE,cpo=TRUE))

summary(Model1)
#Model 2
#space(structured) are modelled through BYM
#time is modelled via random walk (RW1)
#space (unstructured) modelled via iid
formula2 <- y ~ 1 + f(ID.area, model="bym",graph=g) +

```

```
f(ID.year, model="rw1") + f(ID.year1, model="iid")
```

```
Model2 <- inla(formula2, family="poisson", data=dataSC, E=E,
control.predictor=list(compute=TRUE),
```

```
control.compute=list(dic=TRUE, cpo=TRUE))
```

```
summary(Model2)
```

```
#space(area) are modelled through BYM
```

```
#time is modelled via random walk (RW1)
```

```
#space-time interaction is modelled as exchangeable
```

```
formula3 <- y ~ 1 + f(ID.area, model="bym", graph=g) +
```

```
f(ID.year, model="rw1") + f(ID.area.year, model="iid")
```

```
#To obtain the marginal of  $\phi_{ij} + \gamma_{aj}$  we need to create the corresponding  
linear combinations and include these in the model
```

```
lcs = inla.make.lincombs(ID.year = diag(2))
```

```
Model3 <- inla(formula3, family="poisson", data=dataSC, E=E,
```

```
control.predictor=list(compute=TRUE), control.compute=
```

```
list(dic=TRUE, cpo=TRUE),
```

```
lincomb=lcs, control.inla = list(lincomb.derived.only=TRUE))
```

```
#Put the temporal effect ( $\gamma_{aj} + \phi_{ij}$ ) on the natural scale
```

```
summary(Model3)
```

```
#Compute the DIC as a tool for model choice
```

```
Model1$dic$dic
```

```
Model2$dic$dic
```

```
Model3$dic$dic
```

```
#DIC components: Effective number of parameter (pd)
```

```
Model1$dic$p.eff
```

```
Model2$dic$p.eff
```

```

Model3$dic$p.eff
#DIC components: mean.deviance
Model1$dic$mean.deviance
Model2$dic$mean.deviance
Model3$dic$mean.deviance
library(MASS)
library(leaflet)
head(Model3$summary.fitted.values)
RR <- Model3$summary.fitted.values[, "mean"]
RR
cbind(NAME_1,RR)
RR<-cbind(as.factor(NAME_1),round(as.numeric(RR),2))
RR
RRagg<-aggregate(RR[,2],list(RR[,1]), FUN=mean)
RRagg
RRnew<-RRagg$x
RRnew
LL <- Model3$summary.fitted.values[, "0.025quant"]
LL
LL<-cbind(as.factor(NAME_1),round(as.numeric(LL),2))
LLagg<-aggregate(LL[,2],list(LL[,1]), FUN=mean)
LLagg
LLnew<-LLagg$x
LLnew
UL <- Model3$summary.fitted.values[, "0.975quant"]
UL
UL<-cbind(as.factor(NAME_1),round(as.numeric(UL),2))
ULagg<-aggregate(UL[,2],list(UL[,1]), FUN=mean)
ULagg

```

```

ULnew<-ULagg$x
ULnew
pal <- colorNumeric(palette = "YlOrRd", domain = RRnew)
labels <- sprintf("<strong> %s </strong> <br/>
Pop: %s <br/> RRnew: %s (%s, %s)",
Kenya3$NAME_1,
Pop, round(RRnew, 2),
round(LLnew, 2), round(ULnew, 2)
)%>% lapply(htmltools::HTML)
lRR <- leaflet(Kenya3) %>%
addTiles() %>%
addPolygons( color = "grey", weight = 1, fillColor = ~
pal(RRnew),
fillOpacity = 0.5, highlightOptions = highlightOptions
(weight = 4),
label = labels,
labelOptions = labelOptions(
noHide = FALSE,
style =
list(
"font-weight" = "normal", padding = "3px 8px" ),
textsize = "15px", direction = "auto" )
) %>%
addLegend(
pal = pal, values = ~RRnew, opacity = 0.5, title = "RR",
position = "bottomright" )
lRR
#### Probability map
pp<-Spatial.results$pp

```

```

pp
library(leaflet)
pal <- colorNumeric(palette = "YlOrRd", domain = pp)
labels <- sprintf("<strong> %s </strong> <br/>")
Pop: %s <br/> pp: %s ",
Kenya3$NAME_1, Pop, round(pp, 2)
)%>% lapply(htmltools::HTML)
lpp <- leaflet(Kenya3) %>%
addTiles() %>%
addPolygons( color = "grey", weight = 1, fillColor = ~ pal(pp),
fillOpacity = 0.7,
highlightOptions = highlightOptions(weight = 4),
label = labels,
labelOptions = labelOptions(
noHide = FALSE,
style =
list(
"font-weight" = "normal",
padding = "3px 8px" ), fontsize = "15px", direction = "auto"
)
) %>%
addLegend( pal = pal, values = ~pp, opacity = 0.7, title = "pp",
position = "bottomright"
)
lpp

```

Spatial-temporal models for oesophageal cancer cases

Models where smoking is the covariate (oesophageal cancer)

```
library(MASS)
```

```
library(sf)
library(maptools)
library(spdep)
library(raster)
library(plyr)
library(ggplot2)
library(rgdal)
library(SpatialEpi)
library("CARBayesdata")
library("sp")
library(INLA)
library(leaflet)
Kenya<-getData("GADM", country="KE", level=0)
Kenya3<-getData("GADM", country="KE", level=1)
Kenya3
NAME_1<-Kenya3$NAME_1 NAME_1
#plot(Kenya3)
SMdata<-read.csv("D:\\JOSEPH KURIA FOLDER\\PHD 2019 TEX\\
November 2019
\\STOEdataSmoke.csv")
head(SMdata)
Pop<-SMdata$Pop
Pop
Name<-ALdata$NAME_1
Name
Exp2015<-expected(SMdata$Pop,SMdata$Obs2015, 1)
Exp2015
Exp2016<-expected(SMdata$Pop,SMdata$Obs2016, 1)
Exp2016
```

```

SMdata$Exp2015<-expected(SMdata$Pop,SMdata$Obs2015, 1)
SMdata$Exp2015
SMdata$Exp2016<-expected(SMdata$Pop,SMdata$Obs2016, 1)
SMdata$Exp2016
Kenya3 <- merge(Kenya3, SMdata)
Kenya3
low.vector <- as.vector(as.matrix(SMdata[,2:3]))#by column
low.vector
S.vector <- as.vector(as.matrix(SMdata[,5:6]))#by column
S.vector
E.vector <- as.vector(as.matrix(SMdata[,7:8]))#by column
E.vector
year <- numeric(0)
for(i in 1:2){
year<- append(year,rep(i,dim(SMdata)[1]))
}
year
NAME_1<- as.factor(rep(ALdata[,1],2))
NAME_1
SMdata
dataSM<-data.frame(y= low.vector, S=S.vector,E= E.vector,
ID.area=as.numeric
(NAME_1), ID.area1=
as.numeric(NAME_1), year=year, ID.year = year,
ID.year1=year,ID.area.year = seq(1,length(NAME_1)))

dataSM
###Models
nb <- poly2nb(Kenya3)

```



```

nb2INLA("Kenya3.adj", nb)
g <- inla.read.graph(filename = "Kenya3.adj")
### Models with covariates #Model 1 with Space (Structured), Unstructured,
Time and Alcohol use Covariate
formula.1<- y ~ 1+ f(ID.area,model="bym",graph=g)+
f(ID.area1,model="iid")+
f(ID.year,model="rw1")+S
Model.1 <- inla(formula.1,family="poisson",data=
dataSM,E=E, control.predictor=
list(compute=TRUE), control.compute=
list(dic=TRUE,cpo=TRUE))
summary(Model.1)
exp(-7.470)
exp(0.012)
#Model2 with year as the Space (Structured), Unstructured, Time and Alcohol
use Covariate
#global time effect
formula.2<- y ~ 1+ f(ID.area,model="bym",graph=g)+
f(ID.area1,year,model="rw1")
+year+S
Model.2<- inla(formula.2,family="poisson",
data=dataSM,E=E, control.predictor=
list(compute=TRUE), control.compute=
list(dic=TRUE,cpo=TRUE))
summary(Model.2)
#Model.2$fixedeffects
exp( 0.070 )
exp( -7.894 )
exp( 0.045)

```

```

#Non Parametric model alpha + csii + gammaj + phiij #No space time interaction
yet! #csii and are modelled through BYM #gammaj are modelled as RW1
#phiij are modelled as exchangeable
#Space (Structured), Time (Structured), (and) Unstructured, and Alcohol use
Covariate
formula.3<- y ~ 1 + f(ID.area,model="bym",graph=g) +
f(ID.year,model="rw1")+
f(ID.year1,model="iid")+ S
Model.3<- inla(formula.3,family="poisson",data=
dataSM,E=E, control.predictor=
list(compute=TRUE), control.compute=
list(dic=TRUE,cpo=TRUE))
summary(Model.3)
exp(-7.470)
exp(0.012)
#Non Parametric model alpha + csii + gammaj + phiij + deltaij
#csii are modelled through BYM #gammaj are modelled as RW1
#phiij are modelled as exchangeable
#Interaction (deltaij) is modelled as exchangeable
formula.4<- y ~ 1 + f(ID.area,model="bym",graph=g) +
f(ID.year,model="rw1") + f(ID.area.year,model="iid")+S
#To obtain the marginal of phiij + gammaj we need to create the corresponding lin-
ear combinations and include these in the model lcs = inla.make.lincombs(ID.year
= diag(2))
Model.4 <- inla(formula.4,family="poisson",
data=dataSM,E=E, control.predictor=
list(compute=TRUE), control.compute=
list(dic=TRUE,cpo=TRUE), lincomb=lcs,
control.inla = list(lincomb.derived.only=TRUE))

```

```

#Put the temporal effect (gammaj+phij) on the natural scale
summary(Model.4)
exp( -7.033)
exp( 0.034)
#####
#Computethe DIC as a tool for model choice
Model.1$dic$dic
Model.2$dic$dic
Model.3$dic$dic
Model.4$dic$dic
#DIC components: Effective number of parameter (pd)
Model.1$dic$p.eff
Model.2$dic$p.eff
Model.3$dic$p.eff
Model.4$dic$p.eff
#DIC components: mean.deviance
Model.1$mean.deviance
Model.2$dic$mean.deviance
Model.3$dic$mean.deviance
Model.3$dic$mean.deviance
head(Model.4$summary.fitted.values)
#Obtaining Relative Risks, Upper and Lower limits.
RR <- Model4$summary.fitted.values[, "mean"]
RR
RR<-cbind(as.factor(NAME_1),round(as.numeric(RR),2))
RR
RRagg<-aggregate(RR[,2],list(RR[,1]), FUN=mean)
RRagg
RRnew<-RRagg$x

```

```

RRnew
LL <- Model4$summary.fitted.values[, "0.025quant"]
LL
LL<-cbind(as.factor(NAME_1),round(as.numeric(LL),2))
LLagg<-aggregate(LL[,2],list(LL[,1]), FUN=mean)
LLagg
LLnew<-LLagg$x
LLnew
UL <- Model4$summary.fitted.values[, "0.975quant"]
UL
UL<-cbind(as.factor(NAME_1),round(as.numeric(UL),2))
ULagg<-aggregate(UL[,2],list(UL[,1]), FUN=mean)
ULagg
ULnew<-ULagg$x
ULnew
pal <- colorNumeric(palette = "YlOrRd", domain = RRnew)
labels <- sprintf("<strong> %s </strong> <br/>
Pop: %s <br/> RRnew: %s (%s, %s)",
Kenya3$NAME_1,
Pop, round(RRnew, 2),
round(LLnew, 2), round(ULnew, 2)
)%>% lapply(htmltools::HTML)
IRR <- leaflet(Kenya3) %>%
addTiles() %>%
addPolygons(
color = "grey", weight = 1, fillColor = ~ pal(RRnew), fillOpacity = 0.5,
highlightOptions = highlightOptions(weight = 4),
label = labels,
labelOptions = labelOptions(

```

```

noHide = FALSE,
style =
list( "font-weight" = "normal",
padding = "3px 8px" ),
textsize = "15px", direction = "auto" )
) %>%
addLegend(
pal = pal, values = ~RRnew, opacity = 0.5, title = "RR",
position = "bottomright" )
lRR

#### Probability map
pp<-Spatial.results$pp
pp
library(leaflet)
pal <- colorNumeric(palette = "YlOrRd", domain = pp)
pal
labels <- sprintf("<strong> %s </strong> <br/>
Pop: %s <br/> pp: %s ",
Kenya3$NAME_1,
Pop, round(pp, 2)
)%>% lapply(htmltools::HTML)
lpp <- leaflet(Kenya3) %>% addTiles() %>%
addPolygons(
color = "grey", weight = 1, fillColor = ~ pal(pp),
fillOpacity = 0.7,
highlightOptions = highlightOptions(weight = 4),
label = labels,
labelOptions = labelOptions(
noHide = FALSE,

```

```

style =
list(
"font-weight" = "normal",
padding = "3px 8px" ),
textsize = "15px", direction = "auto" )
) %>%
addLegend( pal = pal, values = ~pp, opacity = 0.7, title = "pp",
position = "bottomright" )
lpp

```

Models where alcohol use was the covariate (oesophageal cancer)

```

library(MASS)
library(sf)
library(maptools)
library(spdep)
library(raster)
library(plyr)
library(ggplot2)
library(rgdal)
library(SpatialEpi)
library("CARBayesdata")
library("sp")
library(INLA)
library(leaflet)
Kenya<-getData("GADM", country="KE", level=0)
Kenya3<-getData("GADM", country="KE", level=1)
Kenya3
NAME_1<-Kenya3$NAME_1
NAME_1
#plot(Kenya3)

```

```

ALdata<-read.csv("D:\\JOSEPH KURIA FOLDER\\PHD 2019 TEX\\
November 2019
\\STOEdataAlcohol.csv")
head(ALdata)
Pop<-ALdata$Pop
Pop
Name<-ALdata$NAME_1
Name
Exp2015<-expected(ALdata$Pop,ALdata$Obs2015, 1)
Exp2015
Exp2016<-expected(ALdata$Pop,ALdata$Obs2016, 1)
Exp2016
ALdata$Exp2015<-expected(ALdata$Pop,ALdata$Obs2015, 1)
ALdata$Exp2015
ALdata$Exp2016<-expected(ALdata$Pop,ALdata$Obs2016, 1)
ALdata$Exp2016
Kenya3 <- merge(Kenya3, ALdata)
Kenya3
low.vector <- as.vector(as.matrix(ALdata[,2:3]))#by column
low.vector
A.vector <- as.vector(as.matrix(ALdata[,5:6]))#by column
A.vector
E.vector <- as.vector(as.matrix(ALdata[,7:8]))#by column
E.vector
year <- numeric(0) for(i in 1:2){ year<- append(year,
rep(i,dim(ALdata)[1])) }
year
NAME_1<- as.factor(rep(ALdata[,1],2)) NAME_1 ALdata
dataAL<- data.frame(y= low.vector, A=A.vector,E= E.vector,

```

```

ID.area=
as.numeric(NAME_1),
ID.area1=as.numeric(NAME_1), year=year, ID.year = year,
ID.year1=year, ID.area.year = seq(1,length(NAME_1)))
dataAL
dataAL$SIR<-dataAL$y/dataAL$E
dataAL$SIR
dataAL
###Models
nb <- poly2nb(Kenya3)
nb2INLA("Kenya3.adj", nb)
g <- inla.read.graph(filename = "Kenya3.adj")
### Models with covariates #Model 1 with Space (Structured), Unstructured,
Time and Alcohol use Covariate
formula.1<- y ~ 1+ f(ID.area,model="bym",graph=g)+
f(ID.area1,model="iid")+
f(ID.year,model="rw1")+A
Model.1 <- inla(formula.1,family="poisson",data=dataAL,E=E,

control.predictor=list(compute=TRUE),

control.compute=list(dic=TRUE,cpo=TRUE))
summary(Model.1)
exp(-6.997)
exp(0.034)
#Model2 with year as the Space (Structured), Unstructured, Time and Alcohol
use Covariate #global time effect
formula.2A<- y ~ 1+ f(ID.area,model="bym",graph=g)+
f(ID.area1,year,model="rw1")

```



```

+year+A
Model.2A <- inla(formula.2A,family="poisson",
data=dataAL,E=E, control.predictor=
list(compute=TRUE), control.compute=
list(dic=TRUE,cpo=TRUE))
summary(Model.2A)
#Model.2$fixedeffects
exp( 0.070 )
exp( -7.894 )
exp( 0.045)
#Non Parametric model alpha + csii + gammaj + phij
#No space time interaction yet!
#csii and are modelled through BYM
#gammaj are modelled as RW1
#phij are modelled as exchangeable
#Space (Structured), Time (Structured), (and) Unstructured, and Alcohol use
Covariate
formula.3A<- y ~ 1 + f(ID.area,model="bym",graph=g) + f(ID.year,
model="rw1")+ f(ID.year1,model="iid")+ A
Model.3A <- inla(formula.3A,family="poisson",data=dataAL,E=E,
control.predictor=list(compute=TRUE), control.compute=
list(dic=TRUE,cpo=TRUE))
summary(Model.3A)
#Non Parametric model alpha + csii + gammaj + phij + deltaij
#csii are modelled through BYM #gammaj are modelled as RW1
#phij are modelled as exchangeable

```

```

#Interaction (deltaij) is modelled as exchangeable
formula.4A<- y ~ 1 + f(ID.area,model="bym",graph=g) +

f(ID.year,model="rw1") + f(ID.area.year,model="iid")+A

#To obtain the marginal of phij + gammaj we need to create the corresponding linear combinations and include these in the model lcs = inla.make.lincombs(ID.year = diag(2))
Model.4A <- inla(formula.4A,family="poisson",data=dataAL,E=E,
control.predictor=
list(compute=TRUE), control.compute=
list(dic=TRUE,cpo=TRUE),
lincomb=lcs,control.inla =
list(lincomb.derived.only=TRUE))
#Put the temporal effect (gammaj+phij) on the natural scale
summary(Model.4A)
exp( -7.033)
exp( 0.034)
#####
#Computethe DIC as a tool for model choice
Model.1A$dic$dic
Model.2A$dic$dic
Model.3A$dic$dic
Model.4A$dic$dic
#DIC components: Effective number of parameter (pd)
Model1$dic$p.eff
Model2$dic$p.eff
Model3$dic$p.eff
#DIC components: mean.deviance

```

```

Model1$mean.deviance
Model2$dic$mean.deviance
Model3$dic$mean.deviance
head(Model.1$summary.fitted.values)
#Obtaining Relative Risks, Upper and Lower limits.
RR <- Model4$summary.fitted.values[, "mean"]
RR
RR<-cbind(as.factor(NAME_1),round(as.numeric(RR),2))
RR
RRagg<-aggregate(RR[,2],list(RR[,1]), FUN=mean)
RRagg
RRnew<-RRagg$x
RRnew
LL <- Model4$summary.fitted.values[, "0.025quant"]
LL
LL<-cbind(as.factor(NAME_1),round(as.numeric(LL),2))
LLagg<-aggregate(LL[,2],list(LL[,1]), FUN=mean)
LLagg
LLnew<-LLagg$x
LLnew
UL <- Model4$summary.fitted.values[, "0.975quant"]
UL
UL<-cbind(as.factor(NAME_1),round(as.numeric(UL),2))
ULagg<-aggregate(UL[,2],list(UL[,1]), FUN=mean)
ULagg
ULnew<-ULagg$x
ULnew
pal <- colorNumeric(palette = "YlOrRd", domain = RRnew)
labels <- sprintf("<strong> %s </strong> <br/>")

```

```

Pop: %s <br/> RRnew: %s (%s, %s)",
Kenya3$NAME_1,
Pop, round(RRnew, 2),
round(LLnew, 2), round(ULnew, 2)
)%>% lapply(htmltools::HTML)
lRR <- leaflet(Kenya3) %>%
addTiles() %>%
addPolygons(
color = "grey", weight = 1, fillColor = ~ pal(RRnew), fillOpacity = 0.5,
highlightOptions = highlightOptions(weight = 4),
label = labels,
labelOptions = labelOptions(
noHide = FALSE,
style =
list( "font-weight" = "normal",
padding = "3px 8px" ),
textsize = "15px", direction = "auto" )
) %>%
addLegend(
pal = pal, values = ~RRnew, opacity = 0.5, title = "RR",
position = "bottomright" )
lRR
#### Probability map
pp<-Spatial.results$pp
pp
library(leaflet)
pal <- colorNumeric(palette = "YlOrRd", domain = pp)
pal
labels <- sprintf("<strong> %s </strong> <br/>

```

```

Pop: %s <br/> pp: %s ",
Kenya3$NAME_1,
Pop, round(pp, 2)
)%>% lapply(htmltools::HTML)
lpp <- leaflet(Kenya3) %>% addTiles() %>%
addPolygons(
color = "grey", weight = 1, fillColor = ~ pal(pp),
fillOpacity = 0.7,
highlightOptions = highlightOptions(weight = 4),
label = labels,
labelOptions = labelOptions(
noHide = FALSE,
style =
list(
"font-weight" = "normal",
padding = "3px 8px" ),
textsize = "15px", direction = "auto" )
) %>%
addLegend( pal = pal, values = ~pp, opacity = 0.7, title = "pp",
position = "bottomright" )
lpp

```

Oesophageal cancer models where smoking and alcohol use were covariates

```

library(MASS)
library(sf)
library(maptools)
library(spdep)
library(raster)

```

```

library(plyr)
library(ggplot2)
library(rgdal)
library(SpatialEpi)
library("CARBayesdata")
library("sp")
library(INLA)
library(leaflet)
Kenya<-getData("GADM", country="KE", level=0)
Kenya3<-getData("GADM", country="KE", level=1)
Kenya3
#obtaining the county names
NAME_1<-Kenya3$NAME_1
NAME_1
#loading the data
ALdata<-read.csv("D:\\JOSEPH KURIA FOLDER\\PHD 2019 TEX\\
November
2019 \\STOEdataAlcohol.csv")
ALdata
SMdata<-read.csv("D:\\JOSEPH KURIA FOLDER\\PHD 2019 TEX\\
November 2019\\STOEdataSmoke.csv")
head(SMdata)
Pop<-SMdata$Pop
Pop
Exp2015<-expected(SMdata$Pop,SMdata$Obs2015, 1)
Exp2015
Exp2016<-expected(SMdata$Pop,SMdata$Obs2016, 1)
Exp2016
SMdata$Exp2015<-expected(SMdata$Pop,SMdata$Obs2015, 1)

```

```

SMdata$Exp2015
SMdata$Exp2016<-expected(SMdata$Pop,SMdata$Obs2016, 1)
SMdata$Exp2016
SMdata
Kenya3 <- merge(Kenya3,SMdata)
Kenya3
low.vector <- as.vector(as.matrix(SMdata[,2:3]))#by column
low.vector
S.vector <- as.vector(as.matrix(SMdata[,5:6]))#by column
S.vector
A.vector <- as.vector(as.matrix(ALdata[,5:6]))#by column
A.vector
E.vector <- as.vector(as.matrix(SMdata[,7:8]))#by column
E.vector
year <- numeric(0)
for(i in 1:2){
year<- append(year,rep(i,dim(SMdata)[1]))
}
year
NAME_1<- as.factor(rep(SMdata[,1],2))
NAME_1
dataSM<- data.frame(y= low.vector, S=S.vector, A=A.vector,E= E.vector,

ID.area=as.numeric(NAME_1), ID.area1=as.numeric(NAME_1),

year=year, ID.year = year, ID.year1=year,

ID.area.year =seq(1,length(NAME_1)))

```

```

dataSM
###Models
nb <- poly2nb(Kenya3)
nb2INLA("Kenya3.adj", nb) g <- inla.read.graph(filename = "Kenya3.adj")
### Models with covariates
#Model 1 with Space (Structured), Unstructured, Time and Alcohol use Covariate
formula.1<- y ~1+ f(ID.area,model="bym",graph=g)+ f(ID.area1,
model="iid")+f(year,model="rw1")+A+S
Model.1AS<- inla(formula.1,family="poisson",data=dataSM, E=E,

control.predictor=list(compute=TRUE),

control.compute=list(dic=TRUE,cpo=TRUE))

summary(Model.1AS)
exp(-6.710)#intercept
exp(0.062)#Smoking covariate
exp(-0.018)#Alcohol covariate
#Model2 with year as the Space (Structured), Unstructured, Time and Alcohol
use Covariate
#global time effect
formula.2AS<- y ~1+ f(ID.area,model="bym",graph=g)+
f(ID.area1,year,
model="rw1")+
year+A+S
Model.2AS<- inla(formula.2AS,family="poisson",data=dataSM,E=E,

control.predictor=list(compute=TRUE),

```



```

control.compute=list(dic=TRUE,cpo=TRUE))

summary (Model.2AS)
#Model.2AS$fixed effects
exp(-0.002 )#intercept
exp( -7.737)#year
exp(-0.009)#smoking covariate
exp0.054 )#alcohol covariate
#Non Parametric model alpha + csii + gammaj + phij + deltaij
#csii are modelled through BYM
#gammaj are modelled as RW1
#phij are modelled as exchangeable
#Interaction (deltaij) is modelled as exchangeable
formula.33AS<- y ~ 1 + f(ID.area,model="bym",graph=g) +

f(ID.year,model="rw1") +

f(ID.area.year,model="iid")+A+S+S*A

#To obtain the marginal of phij + gammaj we need to create the corresponding
linear combinations and include these in the model
lcs = inla.make.lincombs(ID.year = diag(2))
Model.3smA<- inla(formula.3smA,family="poisson",data=

dataLSM,E=E, control.predictor=list(compute=TRUE),

control.compute=list(dic=TRUE,cpo=TRUE), lincomb=lcs,

control.inla = list(lincomb.derived.only=TRUE))

```

```

#Put the temporal effect (gammaj+phij) on the natural scale
summary(Model.3AS)
exp(-5.621 )#Intercept coefficient
exp( 0.156)# Smoking covariate coefficient
exp(0.045)#Alcohol use covariate coefficient
exp(-0.003)#Interaction of alcohol use and smoking interaction coefficient
#Computethe DIC as a tool for model choice
Model.1AS$dic$dic
Model.2AS$dic$dic
Model.3AS$dic$dic
#Obtaining relative risks and spatial-temporal maps
RRLSA <- Model.3AS$summary.fitted.values[, "mean"]
RRLSA
RRLSA<-cbind(as.factor(NAME_1),round(as.numeric(RRLSA),2))
RRLSA
RRLSAGagg<-aggregate(RRLSA[,2],list(RRLSA[,1]), FUN=mean)
RRLSAGagg
RRnew<-RRLSAGagg$x
RRnew
cbind(NAME_1,RRnew)
#lower limit relative risks
LLAS<- Model.3AS$summary.fitted.values[, "0.025quant"]
LLAS
LLASG<-cbind(as.factor(NAME_1),round(as.numeric(LLAS),2))
LLASG
LLASGagg<-aggregate(LLASG[,2],list(LLASG[,1]), FUN=mean)
LLASGagg
LLnew<-LLASGagg$x

```

```

LLnew
#upper limit relative risks
ULAS <- Model.3AS$summary.fitted.values[, "0.975quant"]
ULAS
ULASG<-cbind(as.factor(NAME_1),round(as.numeric(ULAS),2))
ULASG
ULASGagg<-aggregate(ULASG[,2],list(ULASG[,1]), FUN=mean)
ULASGagg
ULnew<-ULASGagg$x
ULnew
cbind(NAME_1,ULnew )
# producing spatial-temporal map
pal <- colorNumeric(palette = "YlOrRd", domain = RRnew)
labels <- sprintf("<strong> %s </strong> <br/>
Pop: %s <br/> RRnew: %s (%s, %s)",
Kenya3$NAME_1,
Pop, round(RRnew, 2),
round(LLnew, 2), round(ULnew, 2)
)%>% lapply(htmltools::HTML)
lRRnew <- leaflet(Kenya3) %>%
addTiles() %>%
addPolygons(
color = "grey", weight = 1, fillColor = ~ pal(RRnew),
fillOpacity = 0.5,
highlightOptions = highlightOptions(weight = 4), label = labels,
labelOptions = labelOptions(
noHide = FALSE,
style =
list( "font-weight" = "normal",

```

```
padding = "3px 8px" ),
textsize = "15px", direction = "auto"
)
) %>%
addLegend(
pal = pal, values = ~RRnew, opacity = 0.5, title = "RR",
position = "bottomright" )
lRRnew
```

Spatio-temporal models for lung cancer

Spatio-temporal model for lung cancer where smoking was the covariate

```
library(MASS)
library(sf)
library(maptools)
library(spdep)
library(raster)
library(plyr)
library(ggplot2)
library(rgdal)
library(SpatialEpi)
library("CARBayesdata")
library("sp")
library(INLA)
library(leaflet)
Kenya<-getData("GADM", country="KE", level=0)
Kenya3<-getData("GADM", country="KE", level=1) Kenya3
NAME_1<-Kenya3$NAME_1
NAME_1
```

```

#Reading in the dataset
LSMdata<-read.csv("D:\\JOSEPH KURIA FOLDER\\PHD 2019 TEX\\
November 2019\\LCSM.csv")
head(LSMdata)
Pop1<-LSMdata$Population
Pop1
Name<-LSMdata$NAME_1
Name
Exp2015SM<-expected(LSMdata$Population,LSMdata$Obs2015, 1)
Exp2015SM
Exp2016SM<-expected(LSMdata$Population,LSMdata$Obs2016, 1)
Exp2016SM
LSMdata$Exp2015SM<-expected(LSMdata$Population,LSMdata$Obs2015, 1)
LSMdata$Exp2015SM
LSMdata$Exp2016SM<-expected(LSMdata$Population,LSMdata$Obs2016, 1)
LSMdata$Exp2016SM
Kenya3 <- merge(Kenya3,LSMdata)
Kenya3
LSMdata
low.vector <- as.vector(as.matrix(LSMdata[,2:3]))#by column
low.vector
S.vector <- as.vector(as.matrix(LSMdata[,5:6]))#by column
S.vector
E.vector <- as.vector(as.matrix(LSMdata[,7:8]))#by column
E.vector
year <- numeric(0)
for(i in 1:2){
year<- append(year,rep(i,dim(LSMdata)[1]))
}

```

```

year
NAME_1<- as.factor(rep(LSMdata[,1],2))
NAME_1
LSMdata
dataLSM<- data.frame(y= low.vector, S=S.vector,E= E.vector,

ID.area=as.numeric(NAME_1), ID.area1=as.numeric(NAME_1),

year=year,ID.year = year, ID.year1=year,

ID.area.year = seq(1,length(NAME_1)))

dataLSM
###Models
nb <- poly2nb(Kenya3)
nb2INLA("Kenya3.adj", nb) g <- inla.read.graph(filename = "Kenya3.adj")

### Models with covariates
#Model 1 with Space (Structured), Unstructured, Time and Smoking Covariate
formula.1sm<- y ~1+ f(ID.area,model="bym",graph=g)+

f(ID.area1,model="iid")+f(year,model="rw1")+S

Model.1sm <- inla(formula.1sm,family="poisson",

data=dataLSM,E=E, control.predictor=

list(compute=TRUE),

```

```

control.compute=list(dic=TRUE,cpo=TRUE))

summary(Model.1sm)
exp(-3.421)
exp(0.287 )
#Model2 with year as the Space (Structured), Unstructured, Time and Smoking
Covariate
#global time effect
formula.2sm<- y ~1+ f(ID.area,model="bym",graph=g)+

f(ID.area1,year,model="rw1")+year+S

Model.2sm <- inla(formula.2sm,family="poisson",

data=dataLSM,E=E, control.predictor=

list(compute=TRUE), control.compute=

list(dic=TRUE,cpo=TRUE))

summary(Model.2sm)
#Model.2$fixedeffects
exp( -0.530 )
exp( -2.794 )
exp( 0.182)

#Non Parametric model alpha + csii + gammaj + phij
#No space time interaction yet! #csii and are modelled through BYM #gammaj
are modelled as RW1
#phij are modelled as exchangeable

```

```
#Space (Structured), Time (Structured), (and) Unstructured, and Smiking as
Covariate
```

```
formula.3sm<- y ~ 1 + f(ID.area,model="bym",graph=g) +
```

```
f(ID.year,model="rw1")+ f(ID.year1,model="iid")+ S
```

```
Model.3sm <- inla(formula.3sm,family="poisson",data=
```

```
dataLSM,E=E, control.predictor=list(compute=TRUE),
```

```
control.compute=list(dic=TRUE,cpo=TRUE))
```

```
summary(Model.3sm)
```

```
exp( -3.420 ) exp( 0.288 )
```

```
#Non Parametric model alpha + csii + gammaj + phij + deltaij
```

```
#csii are modelled through BYM
```

```
#gammaj are modelled as RW1
```

```
#phij are modelled as exchangeable
```

```
#Interaction (deltaij) is modelled as exchangeable
```

```
formula.4sm<- y ~ 1 + f(ID.area,model="bym",graph=g) +
```

```
f(ID.year,model="rw1") + f(ID.area.year,model="iid")+S
```

```
#To obtain the marginal of phij + gammaj we need to create the corresponding
linear combinations and include these in the model
```

```
lcs = inla.make.lincombs(ID.year = diag(2))
```

```
Model.4sm <- inla(formula.4sm,family="poisson",
```

```
data=dataLSM,E=E, control.predictor=
```



```

list(compute=TRUE), control.compute=list(dic=TRUE,cpo=TRUE),

lincomb=lcs,control.inla = list(lincomb.derived.only=TRUE))

#Put the temporal effect (gammaj+phiij) on the natural scale
summary(Model.4sm)
exp( -3.371)
exp( 0.338)
Model.4A$mean

#Computethe DIC as a tool for model choice
Model.1sm$dic$dic
Model.2sm$dic$dic
Model.3sm$dic$dic
Model.4sm$dic$dic

#DIC components: Effective number of parameter (pd)
Model.1sm$dic$p.eff
Model.2sm$dic$p.eff
Model.3sm$dic$p.eff
Model.4sm$dic$p.eff

#DIC components: mean.deviance
Model.1sm$mean.deviance
Model.2sm$mean.deviance
Model.3sm$mean.deviance
Model.4sm$mean.deviance

#Obtaining relative risks and spatial-temporal maps
#Obtaining Relative Risks, Upper and Lower limits.
RR <- Model4$summary.fitted.values[, "mean"]
RR

```

```

RR<-cbind(as.factor(NAME_1),round(as.numeric(RR),2))
RR
RRagg<-aggregate(RR[,2],list(RR[,1]), FUN=mean)
RRagg
RRnew<-RRagg$x
RRnew
LL <- Model4$summary.fitted.values[, "0.025quant"]
LL
LL<-cbind(as.factor(NAME_1),round(as.numeric(LL),2))
LLagg<-aggregate(LL[,2],list(LL[,1]), FUN=mean)
LLagg
LLnew<-LLagg$x
LLnew
UL <- Model4$summary.fitted.values[, "0.975quant"]
UL
UL<-cbind(as.factor(NAME_1),round(as.numeric(UL),2))
ULagg<-aggregate(UL[,2],list(UL[,1]), FUN=mean)
ULagg
ULnew<-ULagg$x
ULnew
pal <- colorNumeric(palette = "YlOrRd", domain = RRnew)
labels <- sprintf("<strong> %s </strong> <br/>
Pop: %s <br/> RRnew: %s (%s, %s)",
Kenya3$NAME_1,
Pop, round(RRnew, 2),
round(LLnew, 2), round(ULnew, 2)
)%>% lapply(htmltools::HTML)
IRR <- leaflet(Kenya3) %>%
addTiles() %>%

```

```

addPolygons(
color = "grey", weight = 1, fillColor = ~ pal(RRnew), fillOpacity = 0.5,
highlightOptions = highlightOptions(weight = 4),
label = labels,
labelOptions = labelOptions(
noHide = FALSE,
style =
list( "font-weight" = "normal",
padding = "3px 8px" ),
textsize = "15px", direction = "auto" )
) %>%
addLegend(
pal = pal, values = ~RRnew, opacity = 0.5, title = "RR",
position = "bottomright" )
lRR

#### Probability map
pp<-Spatial.results$pp

pp

library(leaflet)

pal <- colorNumeric(palette = "BuPu", domain = pp)

pal

labels <- sprintf("<strong> %s </strong> <br/>
Pop: %s <br/> pp: %s ",
Kenya3$NAME_1,
Pop, round(pp, 2)
)%>% lapply(htmltools::HTML)

lpp <- leaflet(Kenya3) %>%
addTiles() %>%
addPolygons(

```

```

color = "grey", weight = 1, fillColor = ~ pal(pp),
fillOpacity = 0.7,
highlightOptions = highlightOptions(weight = 4),
label = labels,
labelOptions = labelOptions(
noHide = FALSE,
style = list( "font-weight" = "normal",
padding = "3px 8px"
),
textsize = "15px", direction = "auto"
)
) %>%
addLegend(
pal = pal, values = ~pp, opacity = 0.7, title = "pp",
position = "bottomright" )
lpp

```

Spatio-temporal model for lung cancer where alcohol use was the co- variate

```

library(MASS)
library(sf)
library(maptools)
library(spdep)
library(raster)
library(plyr)
library(ggplot2)
library(rgdal)
library(SpatialEpi)
library("CARBayesdata")

```

```

library("sp")
library(INLA)
library(leaflet)
Kenya<-getData("GADM", country="KE", level=0)
Kenya3<-getData("GADM", country="KE", level=1) Kenya3
NAME_1<-Kenya3$NAME_1
NAME_1
#Reading in the dataset
LALdata<-read.csv("D:\\JOSEPH KURIA FOLDER\\PHD 2019 TEX\\
November 2019\\LCSM.csv")
head(LALdata)
Pop<-LALdata$Population
Pop
Name<-LALdata$NAME_1
Name
Exp2015<-expected(LALdata$Population,LALdata$Obs2015, 1)
Exp2015
Exp2016<-expected(LALdata$Population,LALdata$Obs2016, 1)
Exp2016
LALdata$Exp2015<-expected(LSMdata$Population,LSMdata$Obs2015, 1)
LALdata$Exp2015
LALdata$Exp2016<-expected(LSMdata$Population,LSMdata$Obs2016, 1)
LALdata$Exp2016
Kenya3 <- merge(Kenya3,LALdata)
Kenya3
LALdata
low.vector <- as.vector(as.matrix(LALdata[,2:3]))#by column
low.vector
A.vector <- as.vector(as.matrix(LALdata[,5:6]))#by column

```

```

A.vector
E.vector <- as.vector(as.matrix(LALdata[,7:8]))#by column
E.vector
year <- numeric(0)
for(i in 1:2){
year<- append(year,rep(i,dim(LALdata)[1]))
}
year
NAME_1<- as.factor(rep(LALdata[,1],2))
NAME_1
dataLAL<- data.frame(y= low.vector, A=S.vector,E= E.vector, ID.area=

as.numeric(NAME_1), ID.area1=as.numeric(NAME_1),

year=year, ID.year = year, ID.year1=year,

ID.area.year =seq(1,length(NAME_1)))
dataLAL
###Models
nb <- poly2nb(Kenya3)
nb2INLA("Kenya3.adj", nb) g <- inla.read.graph(filename = "Kenya3.adj")
### Models with covariates
#Model 1 with Space (Structured), Unstructured, Time and Alcohol useCovariate
formula.1AL<- y ~1+ f(ID.area,model="bym",graph=g)+

f(ID.area1,model="iid")+f(year,model="rw1")+A

Model.1AL <- inla(formula.1AL,family="poisson",data=dataLAL,E=E,

```

```

control.predictor=list(compute=TRUE),

control.compute=list(dic=TRUE,cpo=TRUE))

summary(Model.1AL)
exp(-3.500)
exp(0.314 )
#Model2 with year as the Space (Structured), Unstructured, Time andAlcohol
useCovariate
#global time effect
formula.2AL<- y ~ 1+ f(ID.area,model="bym",graph=g)+

f(ID.area1,year,model="rw1")+year+A
Model.2sm <- inla(formula.2AL,family="poisson",data=dataLAL,E=E,

control.predictor=list(compute=TRUE),

control.compute=list(dic=TRUE,cpo=TRUE))

summary(Model.2AL)
#Model.2AL$fixedeffects
exp( -0.455)
exp( -2.822)
exp( 0.167)

#Non Parametric model alpha + csii + gammaj + phij #No space time interac-
tion yet! #csii and are modelled through BYM #gammaj are modelled as RW1
#phij are modelled as exchangeable #Space (Structured), Time (Structured),
(and) Unstructured, and Smiking as Covariate
formula.3AL<- y ~ 1 + f(ID.area,model="bym",graph=g) +

```

```

f(ID.year,model="rw1")+
f(ID.year1,model="iid")+ S
Model.3AL<- inla(formula.3AL,family="poisson",data=dataLAL,E=E,

control.predictor=list(compute=TRUE),

control.compute=list(dic=TRUE,cpo=TRUE))

summary(Model.3AL)
exp( -3.361)
exp( 0.316 )
#Non Parametric model alpha + csii + gammaj + phij + deltaij
#csii are modelled through BYM
#gammaj are modelled as RW1
#phij are modelled as exchangeable
#Interaction (deltaij) is modelled as exchangeable
formula.4AL<- y ~ 1 + f(ID.area,model="bym",graph=g) +

f(ID.year,model="rw1") + f(ID.area.year,model="iid")+A
#To obtain the marginal of phij + gammaj we need to create the corresponding
linear combinations and include these in the model
lcs = inla.make.lincombs(ID.year = diag(2))

Model.4AL <- inla(formula.4sm,family="poisson",data=dataLAL,E=E,
control.predictor=list(compute=TRUE),

control.compute=list(dic=TRUE,cpo=TRUE), lincomb=lcs,

```



```

control.inla = list(lincomb.derived.only=TRUE))

#Put the temporal effect (gammaj+phij) on the natural scale
summary(Model.4AL)
exp( -3.376)
exp( 0.316)
Model.4AL$mean
#Computethe DIC as a tool for model choice
Model.1AL$dic$dic
Model.2AL$dic$dic
Model.3AL$dic$dic
Model.4AL$dic$dic
#DIC components: Effective number of parameter (pd)
Model.1AL$dic$p.eff
Model.2AL$dic$p.eff
Model.3AL$dic$p.eff
Model.4AL$dic$p.eff
#DIC components: mean.deviance
Model.1AL$mean.deviance
Model.2AL$mean.deviance
Model.3AL$mean.deviance
Model.4AL$mean.deviance
#Obtaining Relative Risks, Upper and Lower limits and spatial temporal maps.
RR <- Model4$summary.fitted.values[, "mean"]
RR
RR<-cbind(as.factor(NAME_1),round(as.numeric(RR),2))
RR
RRagg<-aggregate(RR[,2],list(RR[,1]), FUN=mean)
RRagg

```

```

RRnew<-RRagg$x
RRnew
LL <- Model4$summary.fitted.values[, "0.025quant"]
LL
LL<-cbind(as.factor(NAME_1),round(as.numeric(LL),2))
LLagg<-aggregate(LL[,2],list(LL[,1]), FUN=mean)
LLagg
LLnew<-LLagg$x
LLnew
UL <- Model4$summary.fitted.values[, "0.975quant"]
UL
UL<-cbind(as.factor(NAME_1),round(as.numeric(UL),2))
ULagg<-aggregate(UL[,2],list(UL[,1]), FUN=mean)
ULagg
ULnew<-ULagg$x
ULnew
pal <- colorNumeric(palette = "YlOrRd", domain = RRnew)
labels <- sprintf("<strong> %s </strong> <br/>
Pop: %s <br/> RRnew: %s (%s, %s)",
Kenya3$NAME_1,
Pop, round(RRnew, 2),
round(LLnew, 2), round(ULnew, 2)
)%>% lapply(htmltools::HTML)
IRR <- leaflet(Kenya3) %>%
addTiles() %>%
addPolygons(
color = "grey", weight = 1, fillColor = ~ pal(RRnew), fillOpacity = 0.5,
highlightOptions = highlightOptions(weight = 4),
label = labels,

```

```

labelOptions = labelOptions(
  noHide = FALSE,
  style =
  list( "font-weight" = "normal",
  padding = "3px 8px" ),
  textsize = "15px", direction = "auto" )
) %>%
addLegend(
  pal = pal, values = ~RRnew, opacity = 0.5, title = "RR",
  position = "bottomright" )
lRR
#### Probability map
pp<-Spatial.results$pp
pp
library(leaflet)
pal <- colorNumeric(palette = "BuPu", domain = pp)
pal
labels <- sprintf("<strong> %s </strong> <br/>
Pop: %s <br/> pp: %s ",
Kenya3$NAME_1,
Pop, round(pp, 2)
)%>% lapply(htmltools::HTML)
lpp <- leaflet(Kenya3) %>%
addTiles() %>%
addPolygons(
  color = "grey", weight = 1, fillColor = ~ pal(pp),
  fillOpacity = 0.7,
  highlightOptions = highlightOptions(weight = 4),
  label = labels,

```

```

labelOptions = labelOptions(
  noHide = FALSE,
  style = list( "font-weight" = "normal",
  padding = "3px 8px"
),
  textsize = "15px", direction = "auto")
) %>%
addLegend(
  pal = pal, values = ~pp, opacity = 0.7, title = "pp",
  position = "bottomright" )
lpp

```

Lung cancer models where smoking and alcohol use were covariates

```

library(MASS)
library(sf)
library(maptools)
library(spdep)
library(raster)
library(plyr)
library(ggplot2)
library(rgdal)
library(SpatialEpi)
library("CARBayesdata")
library("sp")
library(INLA)
library(leaflet)
Kenya<-getData("GADM", country="KE", level=0)
Kenya3<-getData("GADM", country="KE", level=1)
Kenya3

```

```

#obtaining the county names
NAME_1<-Kenya3$NAME_1
NAME_1
#loading the data
LSMdata<-read.csv("D:\\JOSEPH KURIA FOLDER\\PHD 2019 TEX\\
November 2019\\LCSM.csv")
head(LSMdata)
LALdata<-read.csv("D:\\JOSEPH KURIA FOLDER\\PHD 2019 TEX\\
November 2019\\
LCAL.csv")
head(LALdata)
Pop1<-LSMdata$Population
Pop1
LSMdata$Obs2015
Exp2015SM<-expected(LSMdata$Population,LSMdata$Obs2015, 1)
Exp2015SM Exp2016SM<-expected(LSMdata$Population,LSMdata$Obs2016, 1)
Exp2016SM
LSMdata$Exp2015SM<-expected(LSMdata$Population,LSMdata$Obs2015, 1)
LSMdata$Exp2015SM
LSMdata$Exp2016SM<-expected(LSMdata$Population,LSMdata$Obs2016, 1)
LSMdata$Exp2016SM
Kenya3 <- merge(Kenya3,LALdata)
Kenya3
LALdata
low.vector <- as.vector(as.matrix(LALdata[,2:3]))#by column
low.vector
S.vector <- as.vector(as.matrix(LSMdata[,5:6]))#by column
S.vector
A.vector <- as.vector(as.matrix(LALdata[,5:6]))#by column

```

```

A.vector
E.vector <- as.vector(as.matrix(LALdata[,7:8]))#by column
E.vector
year <- numeric(0)
for(i in 1:2){
year<- append(year,rep(i,dim(LALdata)[1]))
}
year
NAME_1<- as.factor(rep(LALdata[,1],2))
NAME_1
dataLSMA<- data.frame(y= low.vector, S=S.vector, A=A.vector,

E= E.vector, ID.area=as.numeric(NAME_1), ID.area1=

as.numeric(NAME_1), year=year, ID.year = year,

ID.year1=year, ID.area.year =

seq(1,length(NAME_1)))
dataLSMA
###Models
nb <- poly2nb(Kenya3)
nb2INLA("Kenya3.adj", nb) g <- inla.read.graph(filename = "Kenya3.adj")
### Models with covariates
#Model 1 with Space (Structured), Unstructured, Time and Alcohol useCovariate
formula.1sMA<- y ~1+ f(ID.area,model="bym",graph=g)+

f(ID.area1,model="iid")+

f(year,model="rw1")+A+S
Model.1sMA<- inla(formula.1AL,family="poisson",data=

```

```

,E=E,
control.predictor=list(compute=TRUE),

control.compute=list(dic=TRUE,cpo=TRUE))

summary(Model.1smA)
exp(-3.406)#intercept
exp(0.212)#Smoking covariate
exp(0.088)#Alcohol covariate
#Model2 with year as the Space (Structured), Unstructured, Time and Alcohol
use Covariate
#global time effect
formula.2smA<- y ~1+ f(ID.area,model="bym",graph=g)+

f(ID.area1,year,model="rw1")+year+A

Model.2smA <- inla(formula.2smA,family="poisson",data=

dataLSMA,E=E, control.predictor=list(compute=TRUE),

control.compute=list(dic=TRUE,cpo=TRUE))

summaryModel.2smA)
#Model.2smA$fixedeffects
exp( -0.755 )#intercept
exp( -2.583)#year
exp( 0.594)#smoking covariate
exp(-0.380)#alcohol covariate
#Non Parametric model alpha + csii + gammaj + phij + deltaij

```

```

#csii are modelled through BYM
#gammaj are modelled as RW1
#phiij are modelled as exchangeable
#Interaction (deltatij) is modelled as exchangeable
formula.3smA<- y ~ 1 + f(ID.area,model="bym",graph=g) +

f(ID.year,model="rw1") + f(ID.area.year,model="iid")+A+S+S*A
#To obtain the marginal of phiij + gammaj we need to create the corresponding
linear combinations and include these in the model
lcs = inla.make.lincombs(ID.year = diag(2))
Model.3smA<- inla(formula.3smA,family="poisson",

data=dataLSMA,E=E, control.predictor=

list(compute=TRUE), control.compute=

list(dic=TRUE,cpo=TRUE), lincomb=lcs,

control.inla = list(lincomb.derived.only=TRUE))

#Put the temporal effect (gammaj+phiij) on the natural scale
summary(Model.3smA)
exp(-2.342 )#Intercept coefficient
exp( 2.182)# Smoking covariate coefficient
exp(-0.602)#Alcohol use covariate coefficient
exp(-0.075)#Interaction of alcohol use and smoking interaction coefficient
#Computethe DIC as a tool for model choice
Model.1smA$dic$dic
Model.2smA$dic$dic

```



```

Model.3smA$dic$dic
#Obtaining relative risks and spatial-temporal maps
RRLSA <- Model.3smA$summary.fitted.values[, "mean"]
RRLSA
RRLSA<-cbind(as.factor(NAME_1),round(as.numeric(RRLSA),2))
RRLSA
RRLSAGagg<-aggregate(RRLSA[,2],list(RRLSA[,1]), FUN=mean)
RRLSAGagg
RRnew<-RRLSAGagg$x
RRnew
cbind(NAME_1,RRnew)
#lower limit relative risks
LLAS<- Model.3smA$summary.fitted.values[, "0.025quant"]
LLAS
LLASG<-cbind(as.factor(NAME_1),round(as.numeric(LLAS),2))
LLASG
LLASGagg<-aggregate(LLASG[,2],list(LLASG[,1]), FUN=mean)
LLASGagg
LLnew<-LLASGagg$x
LLnew
#upper limit relative risks
ULAS <- Model.3smA$summary.fitted.values[, "0.975quant"]
ULAS
ULASG<-cbind(as.factor(NAME_1),round(as.numeric(ULAS),2))
ULASG
ULASGagg<-aggregate(ULASG[,2],list(ULASG[,1]), FUN=mean)
ULASGagg
ULnew<-ULASGagg$x
ULnew

```

```

cbind(NAME_1,ULnew )
# producing spatial-temporal map
pal <- colorNumeric(palette = "YlOrRd", domain = RRnew)
labels <- sprintf("<strong> %s </strong> <br/>
Pop: %s <br/> RRnew: %s (%s, %s)",
Kenya3$NAME_1,
Pop, round(RRnew, 2),
round(LLnew, 2), round(ULnew, 2)
)%>% lapply(htmltools::HTML)
lRRnew <- leaflet(Kenya3) %>%
addTiles() %>%
addPolygons(
color = "grey", weight = 1, fillColor = ~ pal(RRnew),
fillOpacity = 0.5,
highlightOptions = highlightOptions(weight = 4), label = labels,
labelOptions = labelOptions(
noHide = FALSE,
style =
list( "font-weight" = "normal",
padding = "3px 8px" ),
textsize = "15px", direction = "auto"
)
)%>%
addLegend(
pal = pal, values = ~RRnew, opacity = 0.5, title = "RR", position = "bottom-
right" )
lRRnew

```

APPENDIX B: PUBLICATIONS

Paper I: Poisson-Gamma and Spatial-Temporal Models: with Application to Cervical Cancer in Kenya's Counties

American Journal of Theoretical and Applied Statistics

2021; 10(3): 158-166

<http://www.sciencepublishinggroup.com/j/ajtas>

doi: 10.11648/j.ajtas.20211003.14

ISSN: 2326-8999 (Print); ISSN: 2326-9006 (Online)



Poisson-Gamma and Spatial-Temporal Models: with Application to Cervical Cancer in Kenya's Counties

Joseph Kuria Waitara¹, Gregory Kerich¹, John Kihoro², Anne Korir³

¹School of Sciences and Aerospace Studies, Moi University, Eldoret, Kenya

²School of Computing and Mathematics, The Co-operative University of Kenya, Nairobi, Kenya

³National Cancer Registry, Nairobi, Kenya

Email address:

jkjoseph834@gmail.com (J. K. Waitara), kerichgregoryy@gmail.com (G. Kerich), kihoro.jm@cuk.ac.ke (J. Kihoro),

annkorir@yahoo.com (A. Korir)

To cite this article:

Joseph Kuria Waitara, Gregory Kerich, John Kihoro, Anne Korir. Poisson-Gamma and Spatial-Temporal Models: With Application to Cervical Cancer in Kenya's Counties. *American Journal of Theoretical and Applied Statistics*. Vol. 10, No. 3, 2021, pp. 158-166.

doi: 10.11648/j.ajtas.20211003.14

Received: June 1, 2021; **Accepted:** June 18, 2021; **Published:** June 26, 2021

Abstract: In Africa, Cancer is an emerging health problem where in 2012 new cancer cases were about 847,00 and around 519,00 deaths, three quarters of those deaths occurred in sub-Saharan region. In 2018, cancer was ranked as the third leading cause of deaths in Kenya after infectious and cardiovascular diseases. In 2018 cancer incidences were estimated to be 47,887 new cancer cases and 32,987 deaths. According to data from World Health Organization in 2020, cervical cancer is the second most prevalent cancer among women while breast cancer is the first. In this study, data collected by the Nairobi Cancer Registry (NCR) was used to produce spatial-temporal distribution of the cervical cancer in counties in Kenya. The results showed that counties where data was available among them Embu, Meru, Machakos, Mombasa, Nyeri, Kiambu, Kakamega, Nairobi and Bomet respectively had high risk of cervical cancer. Availability of county-based estimates and spatial-temporal distribution of cervical cancer cases will aide development of targeted county strategies, enhance early detection, promote awareness and implementation of universal coverage of major control interventions which will be crucial in reducing and halting the rising burden of the cancer cases in Kenya. In counties where data was not available the model showed relative risks for cervical cancer disease was minute but it was present, therefore spatial temporal models are very appropriate to estimate relative risks of diseases even when there is a small sample (and possibly without a sample) in a given area by borrowing information from other neighboring regions.

Keywords: Small Area Estimation, Spatial Temporal, Integrated Nested Laplace Approximation, Generalized Linear Mixed Models

1. Introduction

Cancer arises when normal cells transforms into tumour cell in various stages from pre-cancerous lesion to a malignant tumour. According to the International Agency for Research on Cancer (IARC), in 2018 the global new cancer cases was estimated to be 18.1 million and approximately 9.6 million deaths. In Africa, Cancer is an emerging health problem where in 2012 new cancer cases were about 847,00 and 519,00 deaths, three quarters of those deaths occurred in sub-Saharan region. In 2018, cancer was ranked as the third leading cause of deaths in Kenya after infectious and cardiovascular diseases.

The annual incidence of cancer in Kenya was estimated to be 47,887 new cancer cases, with an annual mortality 32,987 in 2018 [6].

Among women cervical cancer is the fourth prevalent cancer worldwide [27]. According to data from World Health Organization in 2020, in Kenya breast cancer in the most prevalent followed by cervical cancer among women [1]. Human papillomavirus (HPV) infection which is transmitted through direct contact is the cause of almost all cervical cancers [9]. According to Schiffman et al. [25] sexually transmitted HPV genotypes, notably HPV16 cause virtually all cervical cancers world-wide if not controlled immunologically or by screening. The control strategies for

cervical cancer includes early screening, vaccination against HPV, treatment of pre-cancerous lesions, diagnosis and treatment of invasive cervical cancer and palliative care [28].

Cervical cancer screening aides in detection of abnormalities which can be treated and pre cancers which may progress into actual cancer thus reducing cervical cancer incidences, deaths and morbidity related to treatment [25].

Spatial temporal components of events associated with time and location can be utilized to reveal aspects related to where and when the events occurred. Cervical cancer is an event associated with time and space (location) therefore analysis can be conducted to determine the trends, spread and patterns to develop ways to halt its spread [20]. This research project presented a spatial temporal approach to analyze spatial dynamics of the cervical cancer cases in Kenya over two years (2015 and 2016) in ten counties namely Bomet, Embu, Kakamega, Kiambu, Machakos, Meru, Mombasa, Nairobi, Nakuru, and Nyeri County.

Generating spatial or spatial-temporal maps by mapping cancer rates may help to identify distribution of cervical cancer in different geographical locations. Identification of

areas with high disease burden helps in prioritization of control efforts and interventions leading to change of risk behaviours [12].

The study main aim was to determine cervical cancer hotspots by density analysis, and evaluate the trend of spatial correlation for the distribution of cases in Kenya from counties with available data.

2. Materials and Methods

The study focused on cervical cancer data collected from different counties for a period of two years (2015 and 2016) by Kenya National Cancer Registry, which is a national Population-Based Cancer Registry (PBCR) that provides high quality cancer surveillance data which includes incidences, mortality, and survival information of cancer patients.

Kenya has area of 582,650 km square and is divided into 47 counties (See Figure 1) and has population of 47.5 Million people as per the census conducted in 2019 by Kenya National Bureau of Statistics (KNBS).



Figure 1. Kenya Administrative Units.

3. Methodology

According to Chandra [7] Small Area Estimation is a twofold problem. First is how to obtain reliable estimates based on data obtained from small areas where some have empty samples and others with very small samples. The second question is how estimation error is to be assessed. Survey data can be used to estimate indicators of small areas with small samples sizes or empty samples. Direct estimators of indicators in small areas might have large sampling errors, their estimation is improved by introducing regression models which provides a relation between independent and dependent variables of interest. Small area estimation is applied in modelling data from this form of non-planned domains (small areas) [2]. Small Area Estimation methods are divided into area and unit level small area models [22].

Gómez-Rubio et al. [11] noted that, fitting Bayesian models using available in-samples data (samples where data is available) can provide reliable estimates for off-samples areas (where data was not available), area level covariates can be included in calculation of the estimates when available. The estimates maybe less accurate as compared when the survey data is available for each area, but they are reasonable since they have lower bias. To model large-scale spatial patterns and with very sparse data, spatial random effects are included at regional level.

In this study the Bayesian Hierarchical Generalized Linear Mixed Models (BHGLMMs) which are used in small area estimation because of their ability to incorporate multiple levels of model dependencies [8] were considered.

Posterior distributions of model parameters are estimated using Bayes method through combination of observed data, prior distributions and the available covariates [13]. Strong spatial autocorrelation is generally exhibited in small area level data

According to Lawson [15], introduction of spatially structured random effects in the model residual spatial autocorrelation is accounted. Conditional autoregressive priors first proposed by Besag et al. [4] are applied in modelling of spatially structured random effects. The equations in the paper were written using Math Type software [19].

3.1. Generalized Linear Mixed Model

In Generalized Linear Mixed Models (GLMMs), the major assumption is that the response variable Y_i comes from an exponential family of the form $Y_i / \theta_i, \phi_i \sim p(\cdot)$.

The exponential family is defined as,

$$p(y_i / \theta_i, \phi_i) = \exp\left(\frac{y_i \theta_i - b(\theta_i)}{a(\phi_i)} + c(y_i, \phi_i)\right) \quad (1)$$

for $i = 1, \dots, n$ observations and θ_i is the scalar canonical parameter.

Equation (2) is obtained by linking the mean $\mu_i = E(y_i / \beta f^i(\cdot)), \phi_i$ through monotonic link function $g(\cdot)$. The link takes different forms depending on the model

applied for example \log for Poisson model.

$$g(\mu_i) = \eta_i = \beta_0 + \sum_{i=1}^n f_i(\mu_i) + \sum_{k=1}^m \beta_k x_{ki} + \varepsilon_i \quad (2)$$

If the form of the functions $f_i(\cdot)$ is varied, it can accommodate different models such as, spatial, spatial-temporal, hierarchical regression and time series. x 's are covariates and ε_i 's are error terms.

3.1.1. Poisson-Gamma Model

A typical Poisson model cannot model extra variance, Poisson-gamma (PG) model is also known as Negative Binomial model which incorporates gamma distributed random-effects is used as an alternative. Poisson-gamma (PG) model, has the following two-level formulation:

$$y_i \sim \text{Poisson}(E_i \theta_i);$$

$$\theta_i \sim \text{Gamma}(\alpha, \beta)$$

In first level the assumption is that the random variable y_i (count data) has Poisson distribution or can be written as $y_i \sim \text{Poisson}(e_i \mu_i \theta_i)$ with probability density function:

$$g(y_i | e_i \mu_i \theta_i) = \frac{e^{-(e_i \mu_i \theta_i)} (e_i \mu_i \theta_i)^{y_i}}{y_i!}, y_i = 0, 1, \dots \quad (3)$$

where $\mu_i = \mu(x_i, \beta)$ is regression model

$x_i = (x_{1i}, x_{2i}, \dots, x_{pi})^T$ is a vector of covariates and

$\beta = (\beta_1, \beta_2, \dots, \beta_p)^T$ are regression coefficients.

In the second stage, θ_i has Gamma distribution or $\theta_i \sim \text{iidGamma}(\alpha, \alpha)$ where the probability density function is:

$$k(\theta_i) = \frac{\alpha^\alpha}{\Gamma(\alpha)} e^{-\alpha \theta_i} \theta_i^{\alpha-1}, \theta_i > 0 \quad (4)$$

based on equation above the joint probability density function is obtained as follows:

$$h(y_i, \theta_i) = \frac{e^{-(e_i \mu_i \theta_i)} (e_i \mu_i \theta_i)^{y_i}}{y_i!} \frac{\alpha^\alpha}{\Gamma(\alpha)} e^{-\alpha \theta_i} \theta_i^{\alpha-1}, y_i = 0, 1, \dots; \theta_i > 0 \quad (5)$$

The marginal probability density function can be obtained as follows

$$m(y_i) = \int_0^\infty h(y_i, \theta_i) d\theta_i$$

$$= \binom{y_i + \alpha - 1}{\alpha - 1} \left(\frac{\alpha}{e_i \mu_i + \alpha} \right)^\alpha \left(1 - \frac{\alpha}{e_i \mu_i + \alpha} \right)^{y_i} \quad (6)$$

The distribution of equation is negative binomial with mean and variance for y_i given as follows:

$$E(Y_i | \underline{\beta}, \alpha) = e_i \mu_i \text{ and } Var(Y_i | \underline{\beta}, \alpha) = e_i \mu_i \left(1 + \frac{e_i \mu_i}{\alpha} \right)$$

The posterior distribution for θ_i is estimated as follows

$$\begin{aligned} \pi(\theta_i | y_i, \underline{\beta}, \alpha) &= \frac{h(y_i, \theta_i)}{m(y_i)} \\ &= \frac{(e_i \mu_i + \alpha)^{y_i + \alpha}}{\Gamma(y_i + \alpha)} e^{-(e_i \mu_i + \alpha)} (\theta_i)^{y_i + \alpha - 1}, \theta_i > 0 \end{aligned} \quad (7)$$

The posterior distribution for θ_i is Gamma based on equation 7 or can be written as:

$$\theta_i | y_i, \underline{\beta}, \alpha \sim \text{Gamma}(y_i + \alpha, e_i \mu_i + \alpha)$$

The Bayes estimate for θ_i , generates the posterior mean and posterior variance a, as follows:

$$\widehat{\theta}_i^B(\underline{\beta}, \alpha) = E_B(\theta_i | y_i, \underline{\beta}, \alpha) = \frac{(y_i + \alpha)}{(e_i \mu_i + \alpha)} \quad (8)$$

and

$$Var_B(\theta_i | y_i, \alpha, v) = \frac{(y_i + \alpha)}{(e_i \mu_i + \alpha)^2} \quad (9)$$

Spatial correlation of the data is not taken in to account in this model since it only introduces a spatially-unstructured over-dispersion [21]. Due to the stated disadvantage and inability to include covariates Poisson-Gamma model has been criticized and has shown to be inferior to Conditional Autoregressive (CAR) convolution models which are more complex [16].

3.1.2. Conditional Autoregressive (CAR) Models

CAR models have been extensively used to model spatial data in various areas such as in epidemiology, agriculture, demography, economy and image analysis as model for both latent and observed variables

To obtain latent Gaussian models, Gaussian priors are assigned to $\beta_0, f_i(\cdot), \beta_k$ and ε_i in equation 2. This can be represented as $\Theta = (\beta_k, s, f_i, s, \dots)$. where Θ is unobserved multivariate Gaussian random variable, whose density $\pi(\Theta / \varphi)$ is controlled by a vector of hyper parameters Φ [17], $\Phi = (\phi_1, \phi_2)$ may not follow Gaussian distribution [18].

The key components of the Latent Gaussian model are, the likelihood of the data $\pi(y / \Theta)$, the Gaussian density of the random vector Θ , $\pi(\Theta / \Phi)$ and the prior distribution of the parameter vector $\pi(\Phi)$.

The posterior is therefore defined as

$$\pi(\Theta, \Phi / y) \propto \pi(\Theta) \pi(\Theta / \Phi) \prod_{i=1}^n \pi(y_i / x_i, \Phi) \quad (10)$$

The posterior marginal's for x_i and posterior marginal's for Φ or some Φ_j can be obtained by applying integrated Nested Laplace approximation (INLA) [17].

3.2. Integrated Nested Laplace Application (INLA) Methodology

This is an appropriate inference based method for approximating the posterior marginal's of the latent Gaussian field $\pi(x_i / y), i = 1, \dots, n$ in three steps.

The posterior marginal's of the latent effects Θ are written as

$$\pi(x_i / y) = \int \pi(x_i / \Phi, y) \pi(\Phi / y) d\Phi \quad (11)$$

$$\pi(\Phi_i / y) = \int \pi(\Phi / y) d\Phi_{-j} \quad (12)$$

The posterior marginal's $\tilde{\pi}(x_i / y)$ and $\tilde{\pi}(\Phi_i / y)$ can be approximated using the Laplace approximation. The first approximation to $\pi(\Phi / y)$ using Gaussian distributions is constructed as follows;

$$\pi(\Phi / y) = \frac{\pi(\Theta, \Phi, y)}{\pi_G(\Theta / y)} \Big|_{\Theta = \Theta^*(\Phi)} \quad (13)$$

where $\tilde{\pi}_G(\Theta / \Phi, y)$ is a Gaussian approximation to the full conditional of Θ and $\Theta^*(\Phi)$ is the mode of the full conditional for, for a given value of Φ . It involves locating the mode of $\tilde{\pi}(\Phi / y)$ which is used to integrate out the uncertainty with respect to Φ when approximating the posterior marginal of x_i .

The posterior marginals of the latent field are supposed to start from $\tilde{\pi}_G(x_i / \Phi, y)$ and approximate the density of $x_i / \Phi, y$ with the Gaussian marginal derived from $\tilde{\pi}_G(\Theta / \Phi, y)$ i.e

$$\tilde{\pi}(x_i / \Phi, y) = N(x_i; (\Phi), \delta_{ii}^2(\Phi)) \quad (14)$$

The marginals of the interest can be computed using numerical integration over a multidimensional grid of values of Φ

$$\tilde{\pi}(x_i / y) = \sum_k \tilde{\pi}(x_i / \Phi_k, y) \tilde{\pi}(\Phi_k / y) \Delta_k \quad (15)$$

where the sum is over the values of Φ with area weights Δ_k [24].

The first step in INLA computation involves approximating the posterior marginal of Φ by using Laplace approximation in equation (13).

The second step involves computing the Laplace

approximation of $\tilde{\pi}(x_i / \Phi, y)$ for selected values of Φ which improves the Gaussian approximation in equation (10)

$$\tilde{\pi}_{LA}(x_i / \Phi, y) \propto \frac{\pi(\Theta, \Phi, y)}{\tilde{\pi}_{GG}(\Theta_{-i} / x_i, \Phi, y)} \Big|_{\Theta_{-i} = \Theta_{-i}^*(x_i, \Theta)} \quad (16)$$

where $\tilde{\pi}_{GG}(\Theta_{-i} / x_i, \Phi, y)$ is a Gaussian approximation to $\Theta_{-i} / x_i, \Phi, y$ around its mode $\Theta_{-i}(x_i, \Phi)$.

An improved version of $\tilde{\pi}_{LA}(x_i / \Phi, y)$ known as Simplified Laplace approximation was developed by Rue et al. [23]. It involves a series of expansion of $\tilde{\pi}_{LA}(x_i / \Phi, y)$ around $x_i = \mu_i(\Phi)$ which corrects for skewness and location and it is also less computationally expensive [23].

The third step involves combining steps 1 and 2 using numerical integration in equation 13 [23].

Cervical Cancer Distribution

Generalized Linear Mixed Model with spatial and temporal random effects and assuming that the response variable was generated by a Poisson process (count data) was used to model cervical cancer cases.

Poisson-gamma (PG) model, has the following two-level formulation:

$$y_i \sim \text{Poisson}(E_i \theta_i);$$

$$\theta_i \sim \text{Gamma}(\alpha, \beta)$$

Where y_i represents observed cervical cancers cases for each county and E_i are the expected cervical cancer cases.

In convolution model, the Poisson regression model is stated as:

$$y_i \sim \text{Poisson}(\mu_i)$$

$$\mu_i = \exp(X_i \beta + \text{offset}_i)$$

Where y_i is the observed cervical cancer cases, X_i 's are the covariates (when available) for the i^{th} observation and offset term is the expected i^{th} population per county.

In this study cervical cancer cases y_i 's were modelled as shown in equation (17) below.

$$g(\mu_i) = \beta_0 + \sum_j \beta_j X_{ij} + f_{trend}(time) + f_{str}(S_i) + f_{unstr}(S_i) \quad (17)$$

Where β_j 's are coefficients, X_{ij} is vector of the covariates, f_{trend} is trend component, f_{str} and f_{unstr} are structured and unstructured spatial effects of the county.

- i. $g(\cdot)$ is a monotonic link function
- ii. β_0 an overall intercept term.
- iii. β_j 's are coefficients
- iv. Correlated random time effects, f_{trend} , to account for time dependence, were modelled via first order random walk [14, 15]. The assumption is values in the

latest year depend on values of the previous year in a specific county. The correlated temporal random effect, f_{trend} , which has a random walk prior distribution, with precision, τ_{ϕ_1} was assigned Gamma (1, 0.001) prior.

- v. The spatial effects $f_{str}(S_i)$, estimated at county level where households were located. The spatial effects by county to account for strong spatial autocorrelation, and were modelled via normal conditionally autoregressive priors (CAR) [4].
- vi. Where $i = 1, \dots, m$, and $j = 1, \dots, T$, are counties and years respectively;

$$u_i | u_j, \tau_u \sim N\left(\frac{1}{\sum_{j=1}^m w_{ij}} \sum_{j \in \delta_i} w_{ij} u_j, \frac{1}{n_{\delta_i} \tau_u}\right) \quad i \neq j \quad (18)$$

where, τ_u is the conditional precision of spatial random effects and δ_i is the neighborhood of the i^{th} region, n_{δ_i} is the

number of neighbours, $\sum_{j=1}^m w_{ij}$, and the spatial weight, w_{ij}

equals 1 if counties i and j are neighbours and zero otherwise. Spatial distribution is utilized to determine the number of neighbours and it is assumed that each county has not less than one neighbour. $\tau_u \sim \text{Gamma}(1, 0.001)$ prior was assigned as the conditional precision for the spatially structured random effect [5].

The Gamma (α, β) density is defined as:

$\pi(\tau) = \frac{\beta^\alpha}{(\alpha-1)} \tau^{(\alpha-1)} \exp(-\beta\tau)$, for, $\tau > 0$, where $\alpha > 0$, the shape parameter, and $\beta > 0$, the inverse scale parameter.

$f_{unstr}(S_i)$ un-correlated random effects by county were assigned a Normal prior, $f_{unstr} \sim N(0, \frac{1}{\tau_v})$, with precision, τ_v assigned Gamma (1, 0.001) prior.

3.3. Model Selection Criteria

Deviance Information Criteria

The Deviance Information Criterion (DIC) [26] similar to Akaike Information Criterion (AIC) and is applied in hierarchical Bayesian models. It is defined for improper priors and provides the effective number of parameters in the Bayesian models.

The deviance is

$$D(x, \theta) = -2 \sum_{i \in I} \log \pi(y_i | x_i, \theta) + \text{constant} \quad (19)$$

The model with smaller DIC value [26] was used to calculate the relative risks. The calculated relative risks are subsequently mapped to geographical areas to produce spatial

temporal maps.

4. Results

Data in this study was analyzed using spatial temporal model R-packages. The packages contain functions for Generalized Linear Mixed Model (GLMM) and INLA methodology.

4.1. Spatial Temporal Maps

This section displays the distribution of notified cases, Standard Incidence Ratios (SIR) map, for all counties with notified cervical cancer cases over two years (2015-2016). Where ($SIR_i > 1$) indicates the risk of cervical cancer is higher, equal ($SIR_i = 1$) or ($SIR_i < 1$) lower risk than that which is expected from the standard population.

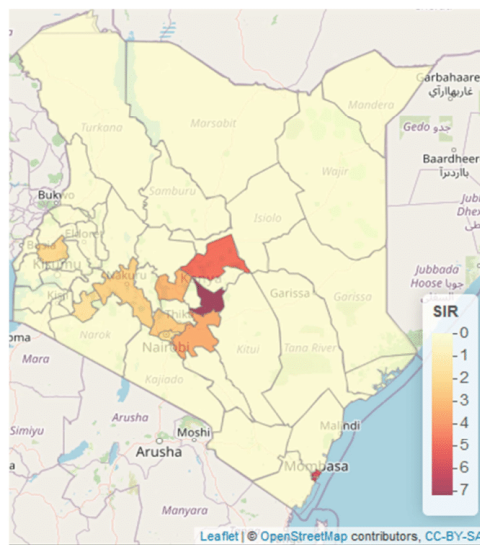


Figure 2. Standardized Incidence Ratios (SIR).

Clearly, from Figure 2 in most counties there was greater risk of cervical cancer cases than expected from the standard population since all counties where data was available had a SIR value greater than 1. Bomet=1.59, Embu =7.13, Kakamega =2.02, Kiambu =2.42, Machakos =3.44, Meru=4.82, Mombasa =5.51, Nairobi=1.66, Nakuru =2.26, Nyeri =3.07.

The deep purple areas indicated higher risks ($SIR > 1$) while the light shaded areas are low risk ($SIR < 1$). The highest burden of cervical cancer cases was in Embu, Mombasa, Meru, Machakos and Nyeri counties respectively.

4.2. Assessing the Presence of Spatial Autocorrelation

Spatial-autocorrelation in spatial-temporal models is modelled via random effects. Therefore, assessment of presence of spatial autocorrelation was conducted by computing the residuals of a fitted simple Poisson log-linear

model.

Table 1. Spatial autocorrelation estimate.

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	8.09e-11	0.0306	4.15	1

Table 2. Over dispersion test.

z-value	p-value	Dispersion parameter
4.15	0.0306	31.202

The over dispersion parameter is equal to 31.202 indicating a relatively high overdispersion, the Poisson model has equal mean and variance assumption.

Moran's I statistic was computed and permutation test for each year of the data to check for spatial-autocorrelation in the model residuals.

The null hypothesis is; there is no spatial autocorrelation while the alternative is; there is positive spatial autocorrelation. The estimated Moran's I statistic was 0.0399 and the p-value was 0.2104 > 0.05, suggesting there was no unexplained spatial autocorrelation in the residuals.

Table 3. Moran I statistic.

Moran I Statistic	p-value
0.0399	0.2104

4.3. Poisson-gamma Model

A Poisson-gamma (Negative-Binomial) model that takes care of over dispersion model and zero inflated variables was fitted. The dispersion parameter for random effect was 1.8692 which was close to 1 and the Akaike Information Criterion (AIC) was 262.9605.

Table 4. Poisson-Gamma estimate.

	Estimate	Std. Error	t-value	Pr(> z)
(Intercept-Fixed effects)	-0.3178	0.3731	-0.852	0.399
Summary of random effects estimates				
Baringo	0.0144	2.5729		
Bomet	2.1385	0.4115		
Bungoma	0.0058	2.5734		

Relative risk estimates for counties where data was available were presented in Table 5. In counties where data was not available the relative risks ranged from 0.01 to 0.06 (Lamu).

Table 5. Relative risks for cervical cancer Poisson-Gamma model.

County	Relative Risk
Bomet	2.14
Embu	9.89
Kakamega	2.78
Kiambu	3.40
Machakos	4.76
Meru	6.48
Mombasa	7.41
Nairobi	2.28
Nakuru	2.19
Nyeri	4.28

The relative risks in Table 5 revealed that Embu county had the highest risk of cervical cancer, followed by Mombasa, Meru, Machakos, Nyeri, Kiambu, Kakamega, Nairobi, Nakuru and Bomet county respectively.

4.4. Spatio-temporal Modelling

Standardized Incidence Ratios (SIRs) sometimes can be useful, but in areas with rare diseases or with small (possibly empty) samples SIRs may be misleading and not very reliable as measure of risk since expected counts may be low as compared to actual observations.

Therefore, estimating disease risks using models which borrow information from neighbouring areas and incorporate covariates information in shrinking or smoothing of extreme values based on small samples is appropriate [10].

The parametric formulation for spatial-temporal models was introduced by Bernardinelli et al. [3], with assumption that the linear predictor can be written as:

$$n_{it} = \alpha + u_i + v_i + \beta t \tag{20}$$

u_i is the spatial structured and v_i unstructured components with a main linear trend β , which represents the overall time effect.

In R-INLA, the first model was specified as follows:
`formula1 <- y ~ 1 + f(Area, model="bym", graph=Kenya.adj) + f(Area1, Year, model="iid") + Year...Model 1`

Area in Model 1 represents the spatial structured effect while Area1 represents spatial unstructured component.

This specification assumes a linear effect of time for each area (i). The parameters estimated by INLA are

$\theta = \{\alpha, \beta, u, v\}$, while the hyper-parameters are $\psi = \{\tau_u, \tau_v\}$.

The assumption of linearity in the i can be realized using a dynamic non parametric formulation for the linear predictor

$$\eta_{it} = \alpha + v_i + v_i + \gamma_t \tag{21}$$

Here, α , v_i and v_i have the same parameterisation as in (20); however, the term γ_t denotes the temporally structured effect, where it is modelled using a first order random walk. In this formulation parameters estimated by INLA are $\theta = \{\alpha, \xi, v, \gamma\}$ and hyper-parameters are represented by $\psi = \{\tau_v, \tau_v, \tau_\gamma\}$.

This second model was specified in R-INLA as:
`formula2 <- y ~ 1 + f(Area, model="bym", graph=Kenya.adj) + f(year, model="rw1") + f(Area1, model="iid")...Model 2`

A third model expanding Model 1 to explain the time trend of cervical cancer cases and to include space-time interaction was specified as follows:

$$\eta_{it} = \alpha + v_i + v_i + \gamma_t + \delta_{it} \tag{22}$$

In this model parameters estimated by INLA are $\theta = \{\alpha, \xi, v, \gamma, \delta\}$ and hyper-parameters are represented by

$$\psi = \{\tau_u, \tau_v, \tau_\gamma, \tau_\delta\}.$$

The third model R-INLA code was formulated as follows:
`formula3 <- y ~ 1 + f(Area, model="bym", graph=Kenya.adj) + f(year, model="rw1") + f(Area1, Year, model="iid")...Model 3`

The DIC values for the three models are presented in Table 6. Despite the added complexity the third model with dynamic time trend and space-time interaction had a smaller DIC value indicating a that the model fitted well to the data and was the most appropriate and the relative risks were obtained from the model.

Table 6. Spatial-temporal models Deviance Information Criterion (DIC) for defined Equations (13)-(15).

Model	\bar{D}	p_D	DIC
Model 1	157.0350	39.2626	196.2976
Model 2	174.4258	30.1654	204.5911
Model 3	158.1371	35.6865	193.8236

The disease risks are usually presented as relative risks in Poisson models. Posterior relative risk distributions greater than 1 indicates an elevated risk of the disease.

Table 7. Relative risks for cervical cancer Poisson-Gamma model.

County	Relative Risk
Bomet	1.53
Embu	7.92
Kakamega	2.06
Kiambu	2.80
Machakos	3.48
Meru	4.43
Mombasa	5.12
Nairobi	1.63
Nakuru	1.95
Nyeri	3.33

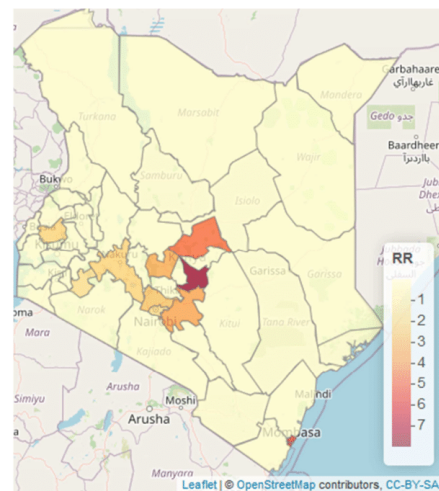


Figure 3. Distribution of the county specific relative risks of cervical cancer in the spatial-temporal model.

The relative risks in Table 7 indicates that Embu county had the highest risk of cervical cancer, followed by Mombasa, Meru, Machakos, Nyeri, Kiambu, Kakamega, Nakuru, Nairobi and Bomet county respectively.

In Figures 3-4 elevated risks are manifested by values greater than 1 in some parts of the country, and a posterior probabilities greater than 0.8.

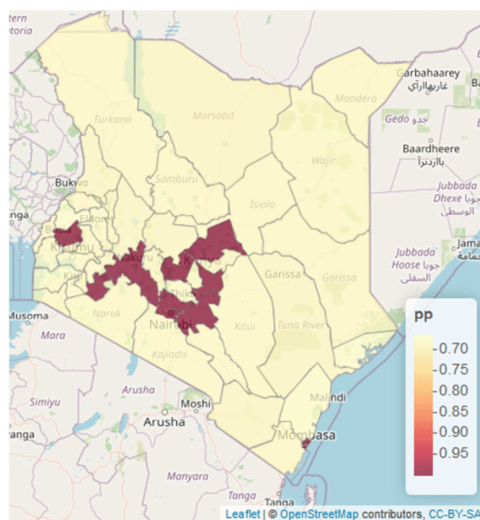


Figure 4. Map of the uncertainty for the spatial effect: $p(\theta_i > 1|y)$.

5. Discussion and Conclusion

The results show that counties where data was available among them Embu, Mombasa, Meru, Machakos and Nyeri counties had very high risk of cervical cancer. The national and county institutions can utilize spatial temporal tools to identify various cancer hot spots and improve screening and treatment facilities based on specific cancer case.

In counties where data was not available the model showed relative risks of cervical cancers was not high but the risk was present, therefore spatial temporal models are very appropriate to estimate relative risks of diseases even when there is a small sample (and possibly an empty sample) in a given area by borrowing information from other neighboring regions. Based on the study findings we recommend the counties with high relatives to create awareness, provide screening services and provide vaccines to the groups which are at higher risk of cervical cancer.

Despite success of this study, the biggest impediment in spatial temporal study is non-availability of adequate county data which will provide more insight on the distribution of cervical cancer cases in Kenya. Therefore the National Cancer Registry in collaboration with counties health departments should work closely to enhance cancer data collection. This will facilitate research and inform the appropriate measures to be implemented in mitigation of the increase of cancer cases.

References

- [1] The Global Cancer Observatory, (2021, March 2021): <https://gco.iarc.fr/today/data/factsheets/populations/404-kenya-fact-sheets.pdf>.
- [2] Roberto Benavent and Domingo Morales. Multivariate fay-herriot models for small area estimation. *Computational Statistics & Data Analysis*, 94: 372–390, 2016.
- [3] L Bernardinelli, D Clayton, C Pascutto, C Montomoli, M Ghislandi, and M Songini. Bayesian analysis of space-time variation in disease risk. *Statistics in medicine*, 14 (21-22): 2433–2443, 1995.
- [4] Julian Besag, Jeremy York, and Annie Mollié. Bayesian image restoration, with two applications in spatial statistics. *Annals of the institute of statistical mathematics*, 43 (1): 1–20, 1991.
- [5] Roger Bivand. *Creating neighbours*. 2019.
- [6] Freddie Bray, Jacques Ferlay, Isabelle Soerjomataram, Rebecca L Siegel, Lindsey A Torre, and Ahmedin Jemal. Global cancer statistics 2018: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: a cancer journal for clinicians*, 68 (6): 394–424, 2018.
- [7] Hukum Chandra. Overview of small area estimation techniques. *Indian Agricultural Statistics Research Institute*, pages 75–88, 2003.
- [8] Avital Cnaan, Nan M Laird, and Peter Slasor. Using the general linear mixed model to analyse unbalanced repeated measures and longitudinal data. *Statistics in medicine*, 16 (20): 2349–2380, 1997.
- [9] Elizabeth TH Fontham, Andrew MD Wolf, Timothy R Church, Ruth Etzioni, Christopher R Flowers, Abbe Herzig, Carmen E Guerra, Kevin C Oeffinger, Ya-Chen Tina Shih, Louise C Walter, et al. Cervical cancer screening for individuals at average risk: 2020 guideline update from the american cancer society. *CA: A Cancer Journal for Clinicians*, 70 (5): 321–346, 2020.
- [10] Alan E Gelfand, Peter Diggle, Peter Guttorp, and Montserrat Fuentes. *Handbook of spatial statistics*. CRC press, 2010.
- [11] Virgilio Gómez-Rubio, Nicky Best, Sylvia Richardson, Guangquan Li, and Philip Clarke. *Bayesian statistics small area estimation*. 2010.
- [12] Marie-Josèphe Horner, Sean F Altekruse, Zhaohui Zou, Louise Wideroff, Hormuzd A Katki, and David G Stinchcomb. Us geographic distribution of prevaccine era cervical cancer screening, incidence, stage, and mortality. *Cancer Epidemiology and Prevention Biomarkers*, 20 (4): 591–599, 2011.
- [13] Diba Khana, Lauren M Rossen, Holly Hedegaard, and Margaret Warner. A bayesian spatial and temporal modeling approach to mapping geographic variation in mortality rates for subnational areas with r-inla. *Journal of data science: JDS*, 16 (1): 147, 2018.
- [14] Leonhard Knorr-Held and G Rasser. Bayesian detection of clusters and discontinuities in disease maps: Simulations. (revised, june 1999). 1999.

- [15] Andrew B Lawson. *Bayesian disease mapping: hierarchical modeling in spatial epidemiology*. Chapman and Hall/CRC, 2013.
- [16] Andrew B Lawson, AB Biggeri, Dankmar Böhning, Emmanuel Lesaffre, Jean-François Viel, ALLAN Clark, PETER Schlattmann, and Fabio Divino. Disease mapping models: an empirical evaluation. disease mapping collaborative group. *Statistics in medicine*, 19 (17): 2217–41, 2000.
- [17] Sara Martino and Håvard Rue. Implementing approximate bayesian inference using integrated nested laplace approximation: A manual for the inla program. *Department of Mathematical Sciences, NTNU, Norway*, 2009.
- [18] Thiago G Martins, Daniel Simpson, Finn Lindgren, and Håvard Rue. Bayesian computing with inla: new features. *Computational Statistics & Data Analysis*, 67: 68–83, 2013.
- [19] Mathtype. <https://www.wiris.com/en/mathtype/office-tools/>. April 2021.
- [20] Kirumba John Mwangi. *Use of GIS in mapping of cancer prevalence a case study of Uasin Gishu County*. PhD thesis, UNIVERSITY OF NAIROBI, 2014.
- [21] Thomas Neyens, Christel Faes, and Geert Molenberghs. A generalized poisson-gamma model for spatially overdispersed data. *Spatial and spatio-temporal epidemiology*, 3 (3): 185–194, 2012.
- [22] JNK Rao. Some new developments in small area estimation. *Journal of the Iranian Statistical Society*, 2 (2): 145–169, 2003.
- [23] Håvard Rue, Sara Martino, and Nicolas Chopin. Approximate bayesian inference for latent gaussian models by using integrated nested laplace approximations. *Journal of the royal statistical society: Series b (statistical methodology)*, 71 (2): 319–392, 2009.
- [24] Håvard Rue and Sara Martino. Approximate bayesian inference for hierarchical gaussian markov random field models. *Journal of statistical planning and inference*, 137 (10): 3177–3192, 2007.
- [25] M Schiffman, J Doorbar, N Wentzensen, S de Sanjose, C Fakhry, BJ Monk, MA Stanley, and S Franceschi. vol. 2. 2016. carcinogenic human papillomavirus infection. *Nature Reviews. Disease Primers*. [Abstract] [Google Scholar], page 16086, 2016.
- [26] David J Spiegelhalter, Nicola G Best, Bradley P Carlin, and Angelika Van Der Linde. Bayesian measures of model complexity and fit. *Journal of the royal statistical society: Series b (statistical methodology)*, 64 (4): 583–639, 2002.
- [27] Chiara Vidoni, Letizia Vallino, Alessandra Ferraresi, Eleonora Secomandi, Amreen Salwa, Menaka Chinthakindi, Alessandra Galetto, Danny N Dhanasekaran, and Ciro Isidoro. Epigenetic control of autophagy in womenâ€™s tumors: role of non-coding mas. *Journal of Cancer Metastasis and Treatment*, 7, 2021.
- [28] WHO (2020, March2020), Human papillomavirus (HPV) and cervical cancer: [https://www.who.int/en/news-room/fact-sheets/detail/human-papillomavirus-\(hpv\)-and-cervical-cancer](https://www.who.int/en/news-room/fact-sheets/detail/human-papillomavirus-(hpv)-and-cervical-cancer).

Paper II: Spatial-temporal Modelling of Oesophageal and Lung Cancers in Kenya's Counties

American Journal of Theoretical and Applied Statistics

2021; 10(4): 175-183

<http://www.sciencepublishinggroup.com/j/ajtas>

doi: 10.11648/j.ajtas.20211004.11

ISSN: 2326-8999 (Print); ISSN: 2326-9006 (Online)



Spatial-temporal Modelling of Oesophageal and Lung Cancers in Kenya's Counties

Joseph Kuria Waitara¹, Gregory Kerich¹, John Kihoro², Anne Korir³

¹School of Sciences and Aerospace Studies, Moi University, Eldoret, Kenya

²School of Computing and Mathematics, The Co-operative University of Kenya, Nairobi, Kenya

³National Cancer Registry, Kenya, Nairobi, Kenya

Email address:

jkjoseph834@gmail.com (J. K. Waitara), kerichgregory@gmail.com (G. Kerich), kihoro.jm@cuk.ac.ke (J. Kihoro),

annkorir@yahoo.com (A. Korir)

To cite this article:

Joseph Kuria Waitara, Gregory Kerich, John Kihoro, Anne Korir: Spatial-temporal Modelling of Oesophageal and Lung Cancers in Kenya's Counties. *American Journal of Theoretical and Applied Statistics*. Vol. 10, No. 4, 2021, pp. 175-183. doi: 10.11648/j.ajtas.20211004.11

Received: June 1, 2021; Accepted: June 18, 2021; Published: June 30, 2021

Abstract: Oesophageal cancer is the cancer that forms in tissues lining the oesophagus (the muscular tube through which food passes from the throat to the stomach) while Lung cancer is the cancer that forms in tissues of the lung, usually in the cells lining air passages. In this study, Data collected by the Nairobi Cancer Registry (NCR) was used to produce spatial-temporal distribution of oesophageal cancer cases for counties in Kenya. The study revealed, counties where data was available Bomet had highest relative risk of oesophageal cancer, followed by Meru, Nyeri, Embu, Nakuru, Kakamega Nairobi, Mombasa, Kiambu and Machakos counties respectively. The study revealed that smoking and alcohol use were significant risk factors of oesophageal cancer in Kenya. Generation of spatio-temporal maps and identification of the risk factors from various counties with notified oesophageal cancer cases is a major milestone since previous studies focused on specific regions. The multiplicative effect of smoking was observed to be 1.012, indicating that oesophageal cancer is 1.2% higher to those who smoke compared to non-smokers. The multiplicative effect of alcohol use was observed to be 1.0346, indicating that oesophageal cancer was 3.5% higher to alcohol users as compared to non-alcohol users. The study findings revealed that, the multiplicative effect of smoking was 1.4021, indicating that lung cancer was 40.21% higher to smokers as compared to non-smokers from the available data. The multiplicative effect of alcohol use was 1.3689 indicating that the risk of lung cancer was 36.89% higher to alcohol users compared to non-alcohol users. Clearly, counties where the data was not available the relative risks were relatively low, therefore even though the data was not available in these counties application of spatial-temporal accounting for covariates revealed that there is risk of oesophageal and lung cancer in the counties. To enhance research on oesophageal, lung and other types of cancer in Kenya the National Cancer Registry in collaboration with Counties health departments should work very closely to enhance cancer data collection to facilitate research and to inform the appropriate measures to be implemented to mitigate the increase of cancer cases.

Keywords: Spatial-temporal, Integrated Nested Laplace Approximation, Generalized Linear Mixed Models

1. Introduction

Oesophageal cancer is the cancer that forms in tissues lining the esophagus (the muscular tube through which food passes from the throat to the stomach) [2]. According to study findings by Schaafsma et al. [18], Ferlay et al. [7] the rate of oesophageal cancer in Kenya is 17.6 per 100,000 which is one of the highest incidence in the Africa continent and is the most common male cancer in Eldoret. Hospital-

based studies conducted in Tenwek hospital in western Kenya by Tenge et al. [20] revealed that male: female ratio of 1.6:1.12 indicating higher incidence rates among males than females. Parts of East Africa and Southern Africa has high burden oesophageal cancers but the risks factors are not fully understood. In South Africa, Tobacco and alcohol have been shown to be clear risk factors [16] but they may not outrightly explain the high rates in East Africa [10]. Kenya is one of a few countries that lie on Africa's oesophageal cancer corridor, which is a region situated in the geographic area of

the Eastern and Western rift-valley and is reported to have the highest incidences in Africa [18]. Therefore a study on the risk factors such as smoking and alcohol use on oesophageal cancer will be very appropriate.

In Kenya prostrate, oesophageal and colorectal cancers are the most prevalent among men while breast, cervical and oesophageal cancers are most common among women. Oesophageal cancer contributes 13.2% of cancer mortality which is the highest, cervical is the second contributing 10% of the cancer deaths while breast cancer comes third at 7.7% [6]. Kenya has a few hospitals which treat oesophageal cancer patients, some of which include Kenyatta National Teaching and Referral Hospital, Moi Teaching Referral Hospital, Tenwek Mission Hospital, Kijabe Mission Hospital, M. P. Shah Hospital/ Cancer Care Kenya.

People with oesophageal cancer may experience: difficulty and pain with swallowing, burning in the chest, frequent choking on food and indigestion or heartburn [1].

Identified alcohol drinking, genetic factors, dietary change/food preparation, and consumption of hot food as the main risk factors for oesophageal cancer in Kenya, they noted that there is a need to investigate the causal relationships between these major risk factors and the development of oesophageal cancer in Kenya [15].

Recent studies on oesophageal cancer has focused on specific regions, therefore mapping its rates, identifying the

risk factors as well as locating counties with high rates will help them prioritize control strategies and design ways to modify risk behaviors. Patel *et al.* [17] conducted a study in Moi Teaching and Referral Hospital (MTRH) in Uasin Gishu County where they identified oesophageal cancer as the leading cancer in men.

Lung cancer is the cancer that forms in tissues of the lung, usually in the cells lining air passages. The two main types are small cell lung cancer and non-small cell lung cancer [2]. According to American Cancer Society [3], the main risk factor for lung cancer is smoking resulting 80% of deaths, where the percentage might be higher for small cell lung cancer (SCLC). Other risk factors includes: Exposure to asbestos and radon a radioactive gas. Bandera *et al.* [4] and Korte *et al.* [11] suggested smoking-adjusted association for high alcohol consumption. Clinical manifestation of lung cancer include: coughing, shortness of breath, wheezing, fever and chest pain [8].

Therefore it is appropriate to conduct the study in Kenya to determine whether smoking and alcohol use are risk factors for oesophageal and lung cancers.

The main aim of the study was to create a spatial temporal model to determine whether smoking and alcohol use are contributing factors of oesophageal and lung cancer cases in Kenya's counties.



Figure 1. Kenya Administrative Units (Counties).

2. Materials and Methods

The study sought to create a spatial-temporal model for oesophageal and lung cancer from the counties data for the year 2015 and 2016. The data in this study was obtained from Kenya National Cancer Registry, which is a national Population-Based Cancer Registry (PBCR). Data includes the total number of oesophageal and lung cancer cases from ten counties namely Bomet, Embu, Kakamega, Kiambu, Machakos, Meru, Mombasa, Nairobi, Nakuru, and Nyeri County.

Kenya is divided into 47 administrative units (See Figure 1) and has a population of 47.5 million as per the Kenya Population and Housing Census that was conducted in 2019.

3. Methodology

The hierarchical Bayes statistical models are specified in hierarchical order since they involve multiple levels. The prior distributions and the covariates are combined then applied to estimate posterior distribution via Bayes method [9]. Data obtained from small areas (e.g county level) generally exhibits spatial autocorrelation. According to Lawson [12], introduction of spatially structured random effects and time varying covariates may account for the spatial autocorrelation in the model.

Integrated Nested Laplace Approximation (INLA) method can be used to estimate the posterior distributions of the parameters in the hierarchical Bayesian model by borrowing strength from the regions with available data to obtain smoothed county level estimates even when the data is sparse [9]. Depending on the available variables and data, various latent models among the convolution, besag and random-walk can be implemented using INLA package in R-software. A Generalized Linear Mixed Model (GLMM) which is a hierarchical Bayesian model was explored and applied to generate results in this study as illustrated in the sections below.

3.1. Generalized Linear Mixed Model

In Generalized Linear Mixed Models (GLMMs) the distribution of the response variable Y_i is assumed to belong to an exponential family as shown in equation (1)

$$p(y_i / \theta_i, \phi_i) = \exp\left(\frac{y_i \theta_i - b(\theta_i)}{a(\phi_i)} + c(y_i, \phi_i)\right) \quad (1)$$

for $i = 1, \dots, n$ observations and θ_i is the scalar canonical parameter. Linking the mean $u_i = E(y_i / \beta f^i(\cdot)), \phi_i$ via monotonic function $g(\cdot)$ generates an additive predictor of the form:

$$g(\mu_i) = \eta_i = \beta_0 + \sum_{i=1}^n f_i(\mu_i) + \sum_{k=1}^m \beta_k X_{ki} + \varepsilon_i \quad (2)$$

Wide range of models may be applied such as spatial, spatial-temporal models and Time series when $f_i(\cdot)$ is varied.

3.2. The Model

Suppose that the index $s' \in (1, 2, \dots, S)$ represents the geographically connected regions. Two regions s and s' are neighbors if they share a common boundary.

According to Moraga [14], Standardized Incidence Ratios (SIRs) can be computed to evaluate disease risk.

For area $i, i = 1, \dots, n$, the SIR is obtained as follows:

$$SIR_i = \frac{Y_i}{E_i}$$

Y_i is the observed counts and E_i is the expected counts.

E_i is calculated using indirect standardization as

$$E_i = \sum_{j=1}^m r_j^{(s)} n_j,$$

$r_j^{(s)}$ is the disease rate in stratum j of the standard population, and n_j is the population in stratum j of the specific area.

Where ($SIR_i > 1$) indicates the risk of cervical cancer is higher, equal ($SIR_i = 1$) or ($SIR_i < 1$) lower risk than that which is expected from the standard population.

SIR may give sense of spatial variability in some situation but it may result to very extreme values when very small or empty samples are involved, due to this shortcoming disease models are preferred to obtain relative risk estimates.

In this study the response variable assumed to be generated by a Poisson process, to model the data. A Generalized Linear Mixed Models assuming a Poisson process with spatial structure, unstructured and temporal random effects was considered.

$$y_i \sim \text{Poisson}(\mu_i)$$

$$\mu_i = \exp(X_i \beta + \text{offset}_i) \quad (3)$$

y_i 's are observed cancer cases (counts per county), X_i 's are the covariates and offset term represents population per county while μ_i is the mean of the observations. The Generalized Linear Mixed Models (GLMM) used to describe the cancer cases y_i is of the form:

$$g(\mu_i) = \beta_0 + \sum_j \beta_j X_{ij} + f_{trend}(time) + f_{str}(S_i) + f_{unstr}(S_i) \quad (4)$$

- i. $g(\cdot)$ is a monotonic link function in our case the \log .
- ii. β_0 is the overall intercept term.
- iii. $\beta_j X_{ij}$, where X_{ij} 's are the covariates β_j 's are the coefficients. The β_j 's for fixed effects ($\beta_j X'_{ij}$) were

assigned normal priors $\beta \sim N(0,100)$. In our model β_j 's are the coefficients of the proportion of smokers and alcohol users of covariates.

- iv. Correlated random time effects, f_{trend} , to account for time dependence, was modelled via first order random walk with precision τ_{ϕ_1} ; assigned $\tau_{\phi_1} \sim \text{Gamma}(1, 0.001)$ prior.
- v. The spatial effects, $f_{str}(S_i)$ estimated at county level. Spatial effects were modelled via normal conditionally autoregressive priors (CAR)[5] to account for spatial auto correlation, the neighbouring counties were assigned weight of 1 and 0 otherwise..
- vi. Non-spatial random effects $f_{unstr}(S_i)$ by county, to model un correlated spatial random effects which was assigned a Normal prior, $f_{unstr} \sim N(0, \frac{1}{\tau_v})$, with precision $\tau_v \sim \text{Gamma}(1, 0.001)$.

The relative risk was presented as μ_i : ($\mu_i > 1$) indicated higher disease risk, ($\mu_i < 1$) lower risk while ($\mu_i = 1$) no risk.

3.3. Model Selection Criteria

Deviance Information Criteria

The Deviance Information Criterion (DIC) [19] is designed for hierarchical models and (in most cases) is well defined for improper priors, it also provides effective number of parameters. The deviance is

$$D(x, \theta) = -2 \sum_{i \in I} \log \pi(y_i/x_i, \theta) + \text{constant} \quad (5)$$

Models fitted was explored to determine contribution of different components namely spatial correlated, uncorrelated random effects, temporal or interactions and the covariates to examine spatial variation in county level oesophageal cancer and lung cancer rates. DIC is based on the deviance of the model penalised for model complexity and its interpretation is similar to the Akaike Information Criterion (AIC), with models having smaller DIC being preferred [19].

4. Results

Data in this study was analyzed using spatial temporal model R-packages. The package contains functions for Generalized Linear Mixed Model (GLMM) and INLA methodology.

4.1. Descriptive Statistics for Oesophageal and Lung Cancers

Table 1. Distribution of oesophageal cancer in 2015.

Gender	Count of Gender	Percentage
Female	349	44.52
Male	435	55.48
Grand Total	784	100

According to data in Table 1, 435 (55.48%) of oesophageal cancer cases were male while 349 (44.52%) of the cases were female.

Table 2. Distribution of oesophageal cancer in 2016.

Gender	Count of Gender	Percentage
Female	289	35.46
Male	526	64.54
Grand Total	815	100

In 2016 as shown in Table 2, 526 (64.54%) of oesophageal cancer cases were male while 289 (35.46%) of the cases were female.

Table 3. Distribution of lung cancer by gender in 2015.

Gender	Count of Gender	Percentage
Female	48	43.24
Male	63	56.74
Grand Total	111	100

In 2015 as shown in Table 3, 63 (56.74%) of lung cancer cases were male while 48 (43.24%) of the cases were female.

Table 4. Distribution of lung cancer by gender in 2016.

Gender	Count of Gender	Percentage
Female	63	43.15
Male	83	56.85
Grand Total	146	100

According to the data in Table 4, in 2016 83 (56.85%) of lung cancer cases were male while 63 (43.15%) of the cases were female.

4.2. Standard Incidence Rates (SIR)

Standard Incidence Rates (SIR) were generated as shown in Figure 2 ($SIR_i > 1$) indicates that area i has higher,

($SIR_i = 1$) equal or lower ($SIR_i < 1$) risk than expected from the standard population. The darker the colour the higher the risk.

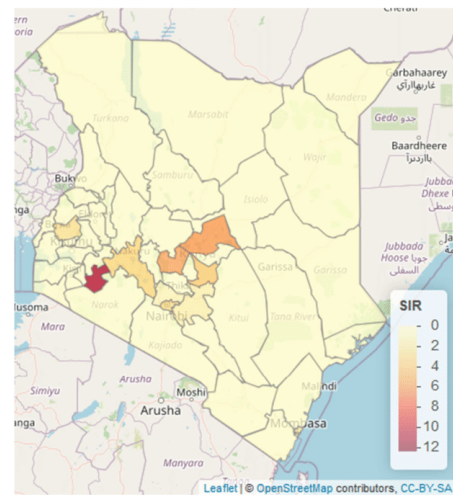


Figure 2. Standardized Incidence Rates (SIR) for oesophageal cancer.

From Figure 2 clearly in most counties there is greater risk of oesophageal cancer cases than expected from the standard population since all counties where data was available has a SIR value greater than 1 except in Kiambu.

Table 5. Standardized Incidence Ratios (SIR).

County	SIR
Bomet	10.09
Embu	4.25
Kakamega	1.91
Kiambu	0.87
Machakos	1.39
Meru	4.22
Mombasa	1.09
Nairobi	2.4
Nakuru	3.08
Nyeri	6.34

4.3. Spatio-temporal Models for Oesophageal Cancer

Four models were fitted, thereafter the most plausible model was selected based on the smallest value of Deviance information Criterion (DIC).

4.3.1. Models Where Smoking Is the Covariate

Model 1: With structured, unstructured spatial effect, trend effects and covariate

In R-INLA the model was specified through the formula as follows:

Model 1 $<-y \sim 1 + f(\text{Counties}, \text{model} = \text{"bym"}, \text{graph} = \text{Kenya.adj}) + f(\text{Counties.1}, \text{model} = \text{"iid"}) + f(\text{Time}, \text{model} = \text{"rw1"}) + \text{smoking}$

Model 2: With structured spatial effect, structured trend effect, global time effect and a covariate

This model was specified as follows:

Model 2 $<-y \sim 1 + f(\text{Counties}, \text{model} = \text{"bym"}, \text{graph} = \text{Kenya.adj}) + f(\text{Counties.1}, \text{model} = \text{"rw1"}) + \text{Time} + \text{smoking}$

Model 3: With structured, unstructured spatial effects, structured trend effects and a covariate

This third model was specified as follows:

Model 3 $<-y \sim 1 + f(\text{Counties}, \text{model} = \text{"bym"}, \text{graph} = \text{Kenya.adj}) + f(\text{Counties.1}, \text{model} = \text{"iid"}) + f(\text{Time}, \text{model} = \text{"rw1"}) + \text{smoking}$

Model 4: structured spatial effect, structured time effect, space-time interaction effects and a covariate

A fourth model allows for an interaction between space and time was specified as follows:

Model 4 $<-y \sim 1 + f(\text{Counties}, \text{model} = \text{"bym"}, \text{graph} = \text{Kenya.adj}) + f(\text{Time}, \text{model} = \text{"rw1"}) + f(\text{Counties. Time}, \text{model} = \text{"iid"}) + \text{smoking}$

Table 6. Results for various models fitted.

Variables	Model 1	Model 2	Model 3	Model 4
Intercept (e^{β_0})	0.001	0.9578	0.0005	0.0005
Smoking (e^{β_1})	1.0121	1.0523	1.0121	1.0121
Year (e^{β_2})	-	0.0004	-	-
DIC	200.91	46067344	200.89	200.63

Table 6 presents the covariate estimates and DIC

components for the four models: despite the added complexity due interaction between space and time, Model 4 was more plausible since it had the lowest DIC value. Model 4 was utilized in obtaining the relative risks per county as shown Table 7 below.

Table 7. The relative risks for counties with notified oesophageal cancer cases where smoking was the covariate.

County	Relative Risk
Bomet	11.71
Embu	2.91
Kakamega	2.28
Kiambu	0.68
Machakos	0.99
Meru	6.68
Mombasa	1.09
Nairobi	1.78
Nakuru	2.59
Nyeri	4.01

The multiplicative effect of smoking was observed to be $e^{\beta_1} = 1.012$, indicating that esophageal cancer is 1.2% higher to those who smoke compared to non-smokers.

4.3.2. Models Where Alcohol Use Is the Covariate

In this section, four models were fitted as in section 4.3.1 where alcohol use was the covariate.

Table 8. Results for various models fitted.

Variables	Model 1	Model 2	Model 3	Model 4
Intercept (e^{β_0})	0.0009	1.0725	0.0009	0.0009
Alcohol use (e^{β_1})	1.0346	1.0460	1.0346	1.0346
Year (e^{β_2})	-	0.0003	-	-
DIC	182.63	81715841	182.74	182.60

Table 8 presents the covariate estimates and DIC components for the four models: despite the added complexity due interaction between space and time, Model 4 was more plausible since it had the lowest DIC value.

The multiplicative effect of alcohol use was observed to be $e^{\beta_1} = 1.0346$, indicating that oesophageal cancer is 3.5% higher to alcohol users as compared to non-alcohol users. The relative risks were obtained for this model as shown in Table 7.

Table 9. The relative risks for counties with notified oesophageal cancer cases with alcohol as the covariate.

County	Relative Risk
Bomet	11.75
Embu	2.80
Kakamega	2.43
Kiambu	0.64
Machakos	0.99
Meru	7.78
Mombasa	1.05
Nairobi	1.78
Nakuru	2.39
Nyeri	3.23

4.3.3. Spatio-temporal Maps

Relative risks for the spatial-temporal distribution are displayed in Figures 3-6. Counties with relative risk greater than 1 had higher risk while those with value of less than 1 had lower risk than expected risk from a standard population. Darker regions indicated relative risk was greater than 1 and while purple colored areas indicated a posterior probability of above 0.8.

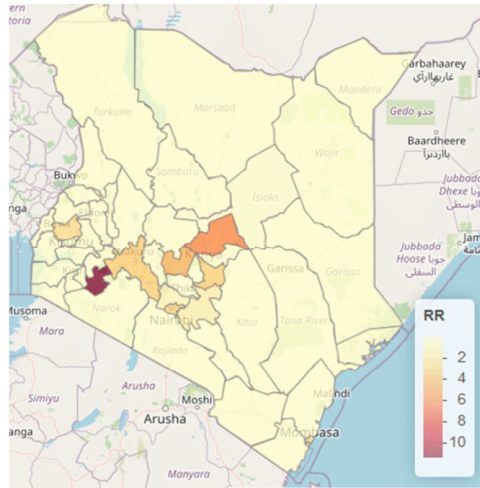


Figure 3. Spatio-temporal distribution of the relative risks for oesophageal cancer with smoking as the covariate.

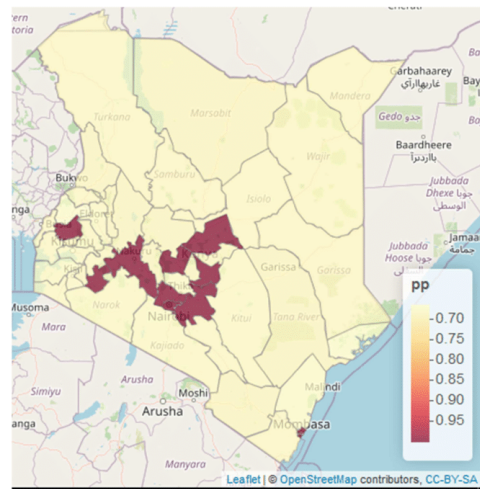


Figure 4. Map of the uncertainty for the spatial temporal effects accounting for smoking effect (oesophageal cancer) $\mu_i; p(\mu_i > 1|y)$.

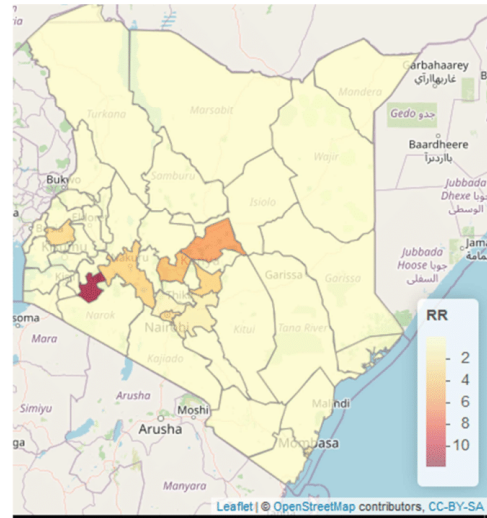


Figure 5. Spatio-temporal distribution of the relative risks for oesophageal cancer with alcohol use as the covariate.

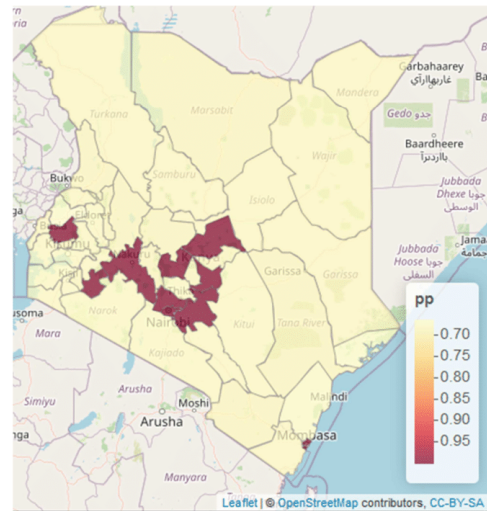


Figure 6. Map of the uncertainty for the spatial temporal effects accounting for alcohol use effect $\mu_i; p(\mu_i > 1|y)$ (oesophageal cancer).

4.4. Spatio-Temporal Models for Lung Cancer

4.4.1. Spatio-Temporal Model for Lung Cancer Where Smoking Was the Covariate

In this section, four models were fitted same as in section 4.3.1 where smoking was the covariate.

Table 10 presents the covariate estimates and DIC components for the four models, Model 4 was selected since it had the lowest DIC value compared to others:

The multiplicative effect of smoking was $e^{\beta_1}=1.4021$, indicating that lung cancer is 40.21% higher to smokers as compared to non-smokers from the available data.

Table 10. Results for various models fitted.

Variables	Model 1	Model 2	Model 3	Model 4
Intercept (e^{β_0})	0.0327	0.5886	0.0327	0.0343
Smoking (e^{β_1})	1.3324	1.1996	1.3338	1.4021
Year (e^{β_2})	-	0.0612	-	-
DIC	129.55	211.78	129.47	127.12

Table 11. The relative risks for counties with notified lung cancer cases with smoking as the covariate.

County	Relative Risk
Bomet	0.68
Embu	5.01
Kakamega	0.19
Kiambu	1.99
Machakos	3.26
Meru	2.42
Mombasa	1.30
Nairobi	3.69
Nakuru	2.02
Nyeri	4.98

Relative risk greater than 1 indicated that the risk of developing lung cancer was higher in the specific counties than in the standard population. The relative risks in Table 11 revealed that majority of the counties where data was available had higher risk of developing lung cancer with exception of Bomet and Kakamega. In Figure 7 the darker the colour the higher the relative risk.

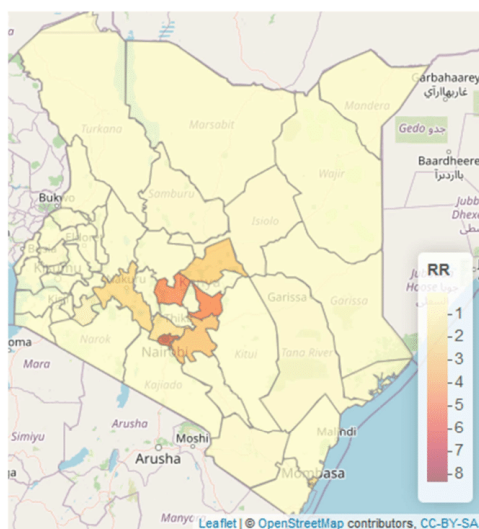


Figure 7. Spatio-temporal distribution of the relative risks for Lung cancer with smoking as the covariate.

4.4.2. Spatio-Temporal Model for Lung Cancer Where Alcohol Use Was the Covariate

Four models were fitted as described in section 4.3.1, where alcohol use was the covariate.

Table 12. Results for various models fitted.

Variables	Model 1	Model 2	Model 3	Model 4
Intercept (e^{β_0})	0.0302	0.6344	0.0347	0.0342
Alcohol use (e^{β_1})	1.3689	0.05948	1.3716	1.3716
Year (e^{β_2})	-	1.1817	-	-
DIC	128.61	209.67	128.77	128.78

Table 12 presents the covariate estimates and DIC components for the four models, Model 1 was selected since it had the lowest DIC value compared to others:

The study findings revealed, the multiplicative effect of alcohol use was $e^{\beta_1}=1.3689$, indicating that the risk of lung cancer is 36.89% higher to alcohol users compared to non-alcohol users.

Table 13. The relative risks for counties with notified lung cancer cases where alcohol use is the covariate.

County	Relative Risk
Bomet	0.69
Embu	5.00
Kakamega	0.19
Kiambu	1.78
Machakos	3.74
Meru	2.54
Mombasa	1.30
Nairobi	4.08
Nakuru	1.80
Nyeri	5.97

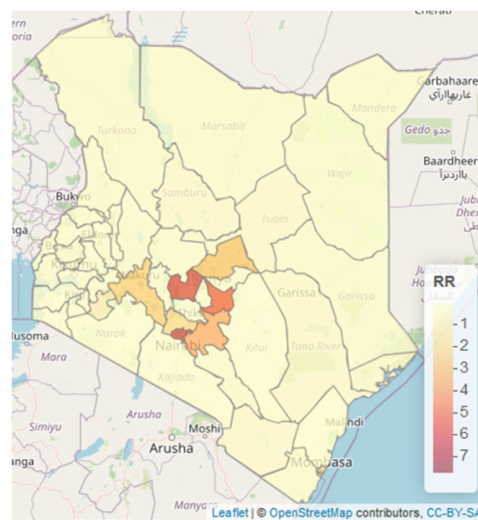


Figure 8. Spatio-temporal distribution of the relative risks for lung cancer with alcohol use as the covariate.

The relative risks in Table 13 indicated that in majority of

the counties where the data was available the risk of developing lung cancer was higher than expected in the standard population. In Figure 8 the darker the colour the higher the relative risk. Nyeri, Embu, Nairobi and Machakos Counties had the highest risks respectively. The relative risk of the areas where the data was not available ranged between 0.0539 and 0.7971.

5. Discussion and Conclusion

The study revealed, counties where data was available Bomet had highest relative risk of oesophageal cancer, followed by Meru, Nyeri, Embu, Nakuru, Kakamega Nairobi, Mombasa, Kiambu and Machakos counties respectively. Other counties had relatively low relative risks which ranged between 0.01-0.08, clearly even though the data was not available in these counties application of spatio-temporal accounting for covariates revealed that there was risk of oesophageal cancer in the counties.

The study revealed that smoking and alcohol use were significant determinants of oesophageal cancer in Kenya. The study findings are consistent with Odera *et al.* [15] who, identified alcohol drinking, genetic factors, dietary change/food preparation, and consumption of hot food as the main risk factors for esophageal cancer. Patel *et al.* [17] showed that there was positive and statistically significant relationship between tobacco smoking and development of oesophageal cancer in Kenya, where in one study smokers had 2.51 odds of developing oesophageal cancer than non-smokers.

Generation of spatio-temporal maps and identification of the risk factors from various counties with notified oesophageal cancer cases is a major milestone since previous studies on oesophageal cancer focused specific regions. Previous studies had indicated that oesophageal cancer is more prevalent in western region of Kenya, but the study revealed that it is also prevalent in other counties such as Meru, Embu and Nyeri.

It is evident that smoking and alcohol use were significant risk factors for lung cancer in Kenya. Meta-analyses conducted by Bandera *et al.* [4] revealed in alcoholics there is risk of lung cancer which is attributable to confounding of residuals since in non-smokers there was no consistent association. Therefore, even though alcohol use is not a direct risk factor for lung cancer it is a confounding risk factor. According to Malhotra *et al.* [13], control of occupational exposures, indoor and outdoor air pollution, understanding the carcinogenic and preventive effects of dietary and other lifestyle factors are some of preventive measures for lung cancer.

The national, county and private health institutions should work closely to create awareness by disseminating information on oesophageal cancer and lung cancer especially in high risk areas as revealed by the study. Screening and treatment facilities should be established based on hot spots of specific cancer cases which are generated from the spatial temporal models.

To enhance research on oesophageal, lung cancer and other types of cancer in Kenya the National Cancer Registry in collaboration with Counties health departments should enhance cancer data collection to facilitate research and to inform the appropriate measures to be implemented to mitigate the increase of cancer cases. We recommend further epidemiological studies to be conducted in areas with high relative risks to find out the other risk factors resulting to higher cases.

References

- [1] Cancer.Net (2021, February 2021), Esophageal Cancer Symptoms and Signs: <https://www.cancer.net/cancer-types/esophageal-cancer/symptoms-and-signs>.
- [2] Gabriela P. (2020, January 2020), Esophageal Cancer, <https://www.webmd.com/cancer/esophageal-cancer>.
- [3] American Cancer Society (2021, February 2021), Lung Cancer risk factors: <https://www.cancer.org/cancer/lung-cancer/causes-risks-prevention/risk-factors.html>.
- [4] Elisa V Bandera, Jo L Freudenheim, and John E Vena. Alcohol consumption and lung cancer: a review of the epidemiologic evidence. *Cancer Epidemiology and Prevention Biomarkers*, 10 (8): 813–821, 2001.
- [5] Julian Besag, Jeremy York, and Annie Mollié. Bayesian image restoration, with two applications in spatial statistics. *Annals of the institute of statistical mathematics*, 43 (1): 1–20, 1991.
- [6] Freddie Bray, Jacques Ferlay, Isabelle Soerjomataram, Rebecca L Siegel, Lindsey A Torre, and Ahmedin Jemal. Global cancer statistics 2018: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: a cancer journal for clinicians*, 68 (6): 394–424, 2018.
- [7] Jacques Ferlay, Isabelle Soerjomataram, Rajesh Dikshit, Sultan Eser, Colin Mathers, Marise Rebelo, Donald Maxwell Parkin, David Forman, and Freddie Bray. Cancer incidence and mortality worldwide: sources, methods and major patterns in globocan 2012. *International journal of cancer*, 136 (5): E359–E386, 2015.
- [8] Dennis Kasper, Anthony Fauci, Stephen Hauser, Dan Longo, J Jameson, and Joseph Loscalzo. *Harrison's principles of internal medicine, 19e*, volume 1. Mcgraw-hill, 2015.
- [9] Diba Khana, Lauren M Rossen, Holly Hedegaard, and Margaret Warner. A bayesian spatial and temporal modeling approach to mapping geographic variation in mortality rates for subnational areas with r-inla. *Journal of data science: JDS*, 16 (1): 147, 2018.
- [10] Anne Korir, Nathan Okerosi, Victor Ronoh, Geoffrey Mutuma, and Max Parkin. Incidence of cancer in n airobi, k enya (2004–2008). *International journal of cancer*, 137 (9): 2053–2059, 2015.
- [11] Jeffrey E Korte, Paul Brennan, S Jane Henley, and Paolo Boffetta. Dose-specific meta-analysis and sensitivity analysis of the relation between alcohol consumption and lung cancer risk. *American journal of epidemiology*, 155 (6): 496–506, 2002.

- [12] Andrew B Lawson. *Bayesian disease mapping: hierarchical modeling in spatial epidemiology*. Chapman and Hall/CRC, 2013.
- [13] Jyoti Malhotra, Matteo Malvezzi, Eva Negri, Carlo La Vecchia, and Paolo Boffetta. Risk factors for lung cancer worldwide. *European Respiratory Journal*, 48 (3): 889–902, 2016.
- [14] Paula Moraga. Small area disease risk estimation and visualization using r. *R J*, 10: 495–506, 2018.
- [15] Joab Otieno Odera, Elizabeth Odera, Jessie Githangâ€™a, Edwin Oloo Walong, Fang Li, Zhaohui Xiong, and Xiaoxin Luke Chen. Esophageal cancer in Kenya. *American journal of digestive disease*, 4 (3): 23, 2017.
- [16] R Pacella-Norman, MI Urban, F Sitas, H Carrara, R Sur, M Hale, P Ruff, M Patel, R Newton, D Bull, et al. Risk factors for oesophageal, lung, oral and laryngeal cancers in black south africans. *British journal of cancer*, 86 (11): 1751–1756, 2002.
- [17] Kirtika Patel, Johnston Wakhisi, Simeon Mining, Ann Mwangi, and Radheka Patel. Esophageal cancer, the topmost cancer at mtrh in the rift valley, kenya, and its potential risk factors. *International Scholarly Research Notices*, 2013, 2013.
- [18] Torin Schaafsma, Jon Wakefield, Rachel Hanisch, Freddie Bray, Joachim Schüz, Edward JM Joy, Michael J Watts, and Valerie McCormack. Africaâ€™s oesophageal cancer corridor: geographic variations in incidence correlate with certain micronutrient deficiencies. *PLoS one*, 10 (10): e0140107, 2015.
- [19] David J Spiegelhalter, Nicola G Best, Bradley P Carlin, and Angelika Van Der Linde. Bayesian measures of model complexity and fit. *Journal of the royal statistical society: Series b (statistical methodology)*, 64 (4): 583–639, 2002.
- [20] CN Tenge, RT Kuremu, NG Buziba, K Patel, and PA Were. Burden and pattern of cancer in western kenya. *East African medical journal*, 86 (1), 2009.