# Spatio-Temporal Modeling of the Kenya Certificate of Primary Education pupil scores through a Bayesian approach

Fredrick Ouma Otiato

I56/68812/2011

A research project submitted in partial fulfillment for the degree of Masters of Science, Social Statistics of the University of Nairobi.

JULY 2013

## DECLARATION

This research project is my original work and has not been presented for examination at any other University.

SIGN: ------------------- DATE ------------------------------

FREDRICK OUMA OTIATO

This project has been submitted for examination with my approval as the University supervisor.

SIGN: ------------------------------ DATE -------------------

PROF. MANENE MOSES M.

## Acknowledgments

I would like to thank the people who have helped me go through this difficult, challenging, yet exciting time of my life. Without their help, this project would not have been successful.

First, I would like to thank God. You have given me the power to believe in myself and pursue my dreams. I could never have done this without the faith I have in You, the Almighty.

Secondly, I cannot express enough gratitude I have for my supervisor, Prof. Manene, who guided me through this project with so much patience. Without your inspiring discussions and patient guidance, I would not have gotten over many hard points of this project

Last and certainly not least, it brings me great pleasure to thank my good friend Erick Arieda Okuto, who gave me excellent advice on this project, especially on the use of the INLA package. I am also grateful to my family for all their support and care. It is your love that supports me and has kept me going in the past, now, and in the future. May God bless you all.

## Executive Summary

This project presents a spatio-temporal model for pupil performance in Kenya certificate Of Primary Education (KCPE) in Kenya between 2006 and 2010 in the 47 counties of Kenya. For this analysis, time will be represented by year (2006-2010) while space will be represented by county. The goal of the project is to put forward an efficient estimation and prediction approach that accounts for both spatial and temporal dependence. The model employs a Bayesian method in which a prior distribution and a likelihood are stated and consequently updated using the data. The model involves a Gaussian Field (GF), affected by a measurement error and a process characterized by time and space.

Data used for this study refers to the KCPE scores of all primary schools in the 47 counties of Kenya from 2006 to 2010. A dependent variable (DV) is created by obtaining aggregate counts of the number of students scoring 350 marks and above in KCPE in each county over the five-year period. Analysis was done using INLA, an R package that makes use of deterministic nested Laplace approximations to provide a faster and more accurate alternative to Markov Chain Monte Carlo (MCMC) methods.

A negative binomial likelihood was assigned to the DV, and, together with a Gaussian prior, space and time attributes were used in a model for explaining performance in KCPE over the five year period and within the 47 counties.

From the analysis, it was found out that throughout the five-year period, the best performance was recorded in 2008. Generally, students in counties located in the central part of the country have the highest probability of scoring at least 350 marks in KCPE while those in the lake and coast regions have the lowest probability. Additionally, performance in any county was seen to be related to that of neighboring counties and the relation became weaker as the distance increases.

## Contents

# CHAPTER 1: INTRODUCTION

## 1.1 Spatial and Spatio-temporal statistics

This chapter will introduce spatial and spatio-temporal statistics in a view to demonstrate the applicability of these methods in this project. Additionally, the chapter will introduce the Bayesian approach to statistics and how it can be incorporated into spatio-temporal models.

Spatial statistics is a branch of Mathematics and includes any analytic approach that studies objects based on their location and distance. This field of study is based on the First Law of Geography as stated by Tobler (1970): “Everything is related to everything else, but near things are more related than distant things”. This law points out to the existence of a positive correlation between closer entities and a weakening of the correlation as the distance between them increases. The law also weakens the independence assumption among observed data. Until recently, there has not been a theory of spatial-temporal processes separate from the already well-established theories of spatial statistics and time series analysis. However, there has been a very rapid growth of research in spatial-temporal data over the last decade due in part to increases in computing power and the availability of spatial data and other spatial tools. Further, there has been an increasing adoption of Bayesian methods in the analysis of such data. Bayesian analysis involves using current information to update our prior knowledge on the parameters of interest.

Spatial data has been categorized into different classes. Cressie (1993) gives three types of such data namely spatial point process, geostatistical, and lattice data.

Spatial point process (sometimes called space-time or spatio-temporal point process) data refers to observations made in a specified study region and at a specified time. Examples of such data include disease incidences, crime rates, or vehicle accidents. The spatial locations are normally captured using variables such as longitude, latitude, and elevation, though sometimes only one or two spatial coordinates are available or of interest. Some other variables such as distance from the sea or some other geographical feature can be used. In this data structure, the locations are sampled from a random process for which we seek inference. Spatial point process data can be used to answer whether observations are similar at all locations and to come up with a predictive model that accounts for spatial and temporal correlation(s) among observed data and provides point estimates for the defined study region.

Geostatistical data refers to observations made from a fixed number of sites. Results from the analysis of such data can then be used to make inference on regions or sites that were not included in the sample. For instance, suppose that rainfall data was made from a number of sites within a country, the geostatistical modeling can be used to predict rainfall quantities in areas where no measurements were made.

The last type of spatial data, known as lattice, is a type of data that consists of aggregate number of counts or measures of a variable of interest for a specified region. Examples include the number of students admitted to secondary school from a given county, or the number of deaths due to road accidents in a given county or district. Analysis of lattice data gives rise to fairly accurate estimates of regions with small sample sizes.

Spatio-temporal modeling becomes important whenever we have data collected across time, say years, months, or days, as well as space. Space refers to the geographical regions or point coordinates of the data while the time element refers to the different time-periods in which data is collected. Consequently, data analysis on spatio-temporal data must account not only for spatial dependence among the covariates, but also recognize the fact that the data forms a time series trend that must be included in the model. In short, the data analysis process must consider spatial and temporal correlations.

In spatial statistics, the following questions are normally of interest;

- Are observations likely to occur in all locations? If not, which locations are more (or less) likely to record the observations.
- Are the covariates more likely to occur in certain locations than others?
- Does an observation either inhibit or promote the occurrence of another variable at another nearby point;
- If observations in one location indeed affect the observations in nearby locations, what is the possible range of influence on neighbors;

Spatial data consists of observations made at distinct points on a surface, rather than a curve, and data is collected in two dimensions as opposed to one dimension as in time series data. It is under this background that we can speak of underlying, latent processes that we would like to model. The inferences which we would like to make relate to these processes, which may or may not be directly observable. For instance, using climate data for example, we could seek to establish the influence of climate on animal migration, or obtain the optimal weather conditions for growth of specific crops, and so on. Indeed, researchers should use observed data to make predictions even in areas that were not included in the sample through an interpolation procedure. Several interpolation techniques have been developed to predict unknown values for any spatial data, such as elevation, rainfall, chemical concentrations in the air, and so forth. A major underpinning of these processes is that observations made in any location are heavily dependent on those made among the neighbors. The inverse distance weighted (IDW) and natural neighbor are some of the most frequently used weighting methods. For this project, a Matern covariance that is a function of the distances between an entity and its neighbors will be used.

## 1.2 Trends in the Kenya Certificate of Primary Education examinations

The Kenya Certificate of Primary Education (KCPE) is an examination that has been administered by the Kenya National Examination Council (KNEC) since 1985 when the current 8-4-4 education systems was launched. Under this system, children attend primary education for a minimum of 8 years, 4 years of secondary education, and 4 years of university education.

Today the KCPE examination is used as an eligibility criterion for joining high school. Previously, the maximum marks possible was 700 based on 7 subjects, however, the system was revised in 2001 and the total examinable subjects was reduced to 5, implying a maximum score of 500 marks. Pupils are examined in 5 subjects namely Maths, English, Swahili, Social Studies, and Religious Studies (Christian/Islamic/Hindu). Previously, analysis of pupil performance in KCPE was done either at district or province level, however, since the establishment of counties in 2009, analyses have changed and are today done at county level.

Studies have shown interesting relationships in performance in KCPE examinations in the 47 counties of Kenya both in individual subjects and in terms of aggregate and average scores. Generally, counties located in the arid and semi-arid lands (ASAL) such as Mandera, Garissa, and Wajir have had poor results while those found in fertile lands such as Elgeyo Marakwet, Nandi, Uasin Gishu, and Kirinyaga have had higher scores. Additionally, the number of persons sitting for the examination has also been low in counties located in ASAL areas as opposed to those in other areas.

The low enrolment rates in arid areas can be attributed to the low population density in the counties. For instance, Mandera, Wajir and Garissa have population densities of 39.47, 11.68 and 14.10 respectively, while Nandi, Uasin Gishu and Elgeyo Marakwet have population densities of 261.07, 267.30  and 122.12 respectively. A low population density implies a smaller percentage of persons enrolling for examination. Another factor that could explain the low enrolment rates in ASAL areas is the low educational aspirations of persons living in these areas. Studies show that persons living in rural areas have low educational aspirations as compared to their urban counterparts (Xu, 2009). Additional studies have found out that rural students place less value on academics and are usually focused more on non-academic qualities. It is known that most parts of arid and semi-arid lands can be considered as rural areas and this leads to low educational aspirations and less value of academics among locals. The net effect of the two factors is a low pupil enrolment rate and low average student scores in the affected counties.

Student performance in KCPE examination has been seen to be affected by poverty index. It is known that strong, secure relationships in early childhood help stabilize a child’s behavior and provides the key ingredients to building lifelong social skills. A child who grows up with such relationships learns healthy, proper emotional responses to daily events. However, children raised in poor households normally fail to acquire these responses and this goes on to affect their performance in school. For instance, pupils who lack emotional stability may so easily get frustrated and give up on classroom tasks.

Apart from emotional stability, poverty levels affect the social economic status of parents. Even with the introduction of free primary education (FPE) in 2003, poor households are finding a challenge as they cannot afford other educational costs such as uniform, stationery, examination fees and other charges levied by the schools. Consequently, children from poor households area not able to compete with other children effectively.  Indeed, it has been shown that counties with high poverty indices have the lowest student performance in KCPE. For instance, according to a recent audit of wealth and poverty in Kenya, Kajiado is the richest county. Only 12% of people in the county are classified as poor. However, in Turkana, which is considered the poorest county according to the same audit, 94% of the residents are considered poor. Ranked in terms of mean average student marks in the 2012 KCPE examination, Kajiado comes 13th while Turkana is 27th (Jagero, 2013).

However, poverty index alone is not sufficient in explaining spatial variations in the performance of counties in KCPE examination as some counties that have low poverty indices are seen to have performed dismally and vice versa. For instance, West Pokot county which ranks 40th out of 47 counties on the poverty index ranks 9th in the KCPE ranking while Lamu, ranking 6th on the poverty index, ranks 42nd in KCPE performance. Consequently, there is need to include other covariates in the model such as those pertaining to temporal phenomena.

There have been changing patterns of enrolment for KCPE examination in Kenya for the past two decades. One notable change is the progressive increase in enrolment rates because of the introduction of free primary education (FPE) in 2003 as shown below;

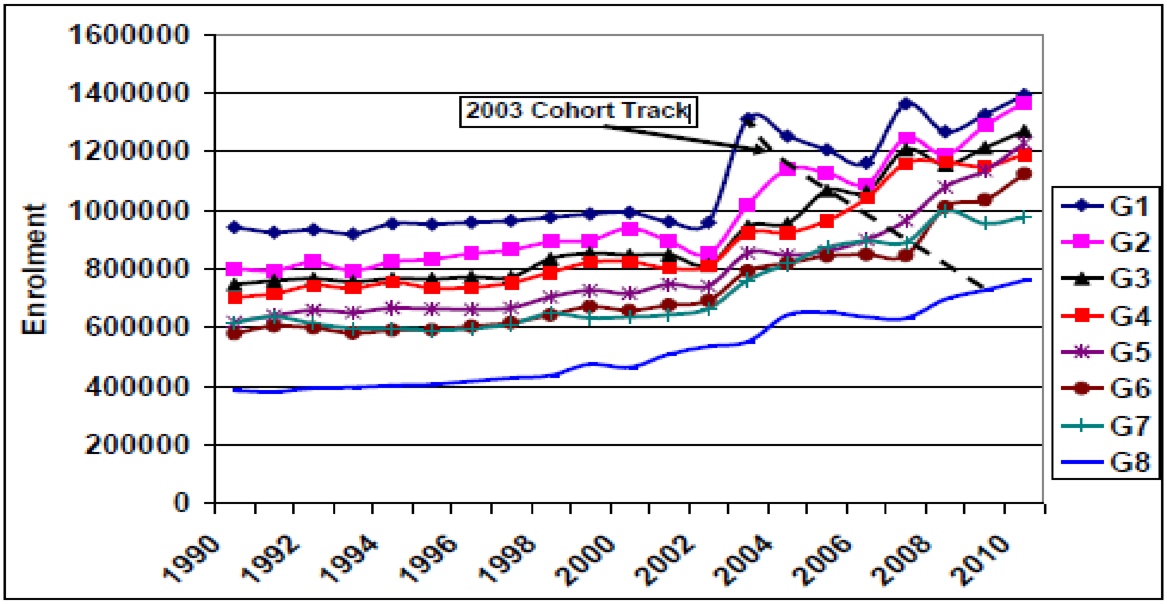

Figure 1: Enrolments by Year and Grade, Kenya

The introduction of FPE has ensured that children from poor families are able to attain primary education and putting the country into the roadmap to attaining one of the milestones of the millennium development goals. Although FPE has led to an increase in school attendance and corresponding enrolment in KCPE examination, it has been noted that some disparities still exist between urban and rural schools, and between children from poor and rich families. Importantly, children from poor families have been seen to be more likely to be over-age by the time they sit for their final examination. Over-age among children from poor families is caused by repetition of grades while some children have stunted growth and thus are enrolled into school late. Repetition of classes is common in areas that have low completion rates and this augments the number of persons not completing primary education. In addition, it has been shown that the more over-age a child is, the more it is likely that they will underachieve in their final examination as shown in the graph below;

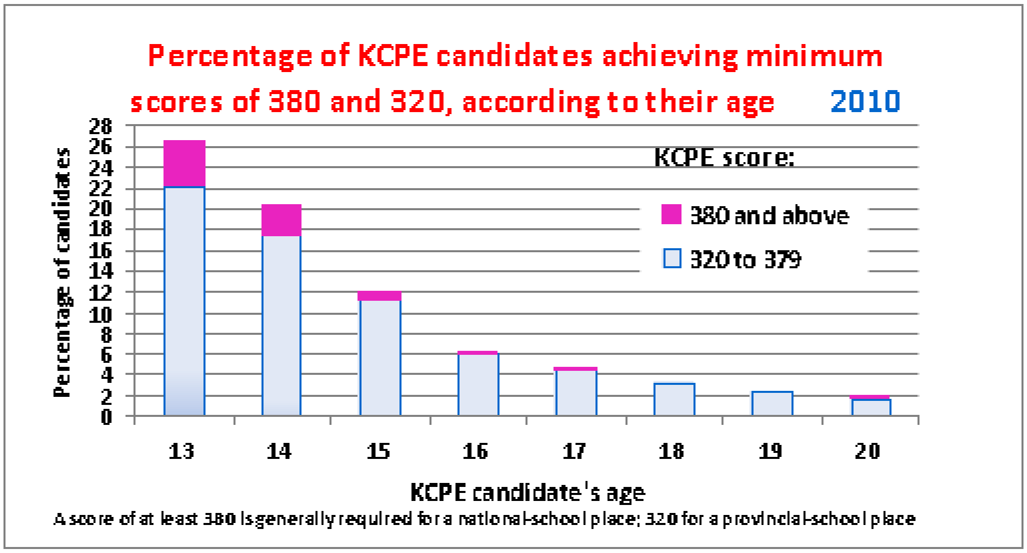

Fig. 2: Percentage of KCPE candidates achieving minimum scores of 380 and 320, by age

Further, where over-age persons share classes with younger children they may have psychological problems such as self-esteem hence affecting their concentration and cognitive abilities.

Apart from being less likely to get good grades, over-age persons are also likely to drop out of school. This stems out of the fact that these persons normally have an average age of between 14 and 15. At this age, most children normally enter the job market. Their chances of continuing with education are thus severely reduced. Over-age among children in any grade is prevalent among poor households, and since most of these families are located in rural areas, enrolment in any grade and the eventual number of persons sitting for KCPE examination is lowest in counties predominantly consisting of rural areas.

In contrast, being over-age can be beneficial to children particularly where a large proportion of children in a class are over-age. Over-age children have better cognitive abilities and hence grasp new concepts faster than younger children. This may lead to over-age children performing better than younger children and the concept may be used to explain anomalies observed in examination performance in which some counties that are expected to perform poorly record unexpectedly better results. For instance, West Pokot county which consists mostly of rural areas ranked number 8 out of 47 counties in the 2012 KCPE exam when it would have been expected to perform much worse.

Generally, performance in the KCPE examination has been observed to differ from county to county and from individual to individual.

## Bayesian Methods

There are two major approaches to statistics in use today: frequentist/classical approach and Bayesian approach.  The former is the most commonly used method. Under the frequentist approach, emphasis is on the probability of the data, given the hypothesis i.e. data is treated as random (different outcome for every study) while the hypothesis is fixed (either true, 1, or false, 0). The word frequentist is used since it is concerned with the frequency with which one expects to observe the data, given some hypothesis about some phenomenon. The *p*-values used by frequentists is usually the expression of the probability of the data given the hypothesis.

Contrary to frequentists' approach, Bayesian inference focuses on the probability of the hypothesis given the data. The approach treats data as fixed (this is the only data available) and the hypotheses as random (the hypothesis might be true or false, with some probability between 0 and 1). Bayes' theorem is central to calculating this probability. While parameters are treated as constants by frequentists, Bayesians treat them as random variables that take on different values which are updated as more evidence becomes available (through data collection).

From a general perspective, the goal of Bayesian statistics is to represent prior uncertainties about model parameters with a probability distribution and to update this prior information with current data that produces a posterior distribution for the parameter that contains less uncertainty. This approach results into a subjective view of statistics as opposed to the classical view. From the Bayesian perspective, any quantity for which the true value is uncertain, including model parameters, can be represented with probability distributions. From the classical perspective, however, parameters are considered as fixed entities and not as probability distributions. Only the data are random, and thus, probability distributions can only be used to represent the data.

The process of repeating the test and updating our posterior probability is the basic concept in Bayesian statistics. From Bayesian thinking, we begin with some prior probability for some event of interest, and then use data collected to obtain a posterior probability which can be used for subsequent analyses. In Bayesian milieu, this is a suitable method for carrying out scientific research, i.e. we continue gathering data to evaluate a specific scientific inquiry rather than begin a new (blind) one each time the query is encountered, because previous research gives us *priori* information concerning the merit of the hypothesis.

## 1.3 Problem Statement

The last two decades have seen an increasing interest in school performance as parents 'shop around' for the best schools for their children, particularly after the introduction of free primary education. Today, it is theoretically possible for parents to choose a good school for their children and this process has been made easier by annual publications of school performance soon after release of national examination results. This choice is normally based on factors such as school size, trends in examination performance over the past few years, school categorization (district, provincial or national), its location (rural/urban), years of existence, and admission points. In practice, however, a parent may not make a good school choice due to latent factors or some unforeseen events. For instance, it is known that good schools normally fill up very fast and unless quality control checks are instituted, they may experience a decline within a few years. A time series model may thus be used to model and predict performance thus help parents and other stakeholders in making informed decisions. However, this aspect is normally overlooked in many analyses. A time series trend also implies that the relationship in examination results observed between any two time-periods becomes weaker as time increases. In addition, from Tobler's law as mentioned above, when a district or county performs well (or poorly) in examination, then its neighbors are also more likely to have more or less similar results. This law points to spatial dependence in school performance among schools which should be used when analyzing school performance, and consequently when choosing a school.

Student performance in examinations, e.g. in terms of aggregate counts of passes and one or more covariates, is frequently characterized by spatial and/or temporal characteristics that need to be taken into consideration in the inferential process. For this reason, the modeling process has to incorporate not only the spatial dependence of the variables of interest but also check for time series trends i.e. one has to account for temporal correlations as well as spatial correlations in the data.

It is known that students differ in gender, culture, religion, language, home environment, financial status of parents etc., whereas the schools differ in size of students, quality of teacher, learning facilities available, location of the school, government policies etc. This project will focus on the hitherto less researched area of spatial dependence in students' performance.

Data on primary school examination obviously exhibit some very important spatial variations or similarities that, if explored further, can add vital information to models used to explain school or regional performance in terms of aggregate counts or mean scores. Despite a considerable quantity of research into factors affecting school performance, only a small proportion of studies have focused on spatial and time dependence in examination performance, and this is the main motivation of this project. This information can help parents in determining the best schools for their children while government institutions and other stakeholders can use the data for planning purposes.

## 1.4 Objective

The objective of this project is to analyze and model the examination scores of pupils in KCPE from 2006 to 2010 in the 47 counties of Kenya based on spatial and temporal characteristics.

Specific Objectives

- Formulate a model accounting for time and space covariates in KCPE examination scores in Kenya
- Compute the probability of a student obtaining threshold marks in KCPE for each county
- To construct a spatial dependence map for KCPE examination scores in the 47 counties of Kenya

## 1.5 Significance of the Study

This study aims to build a spatio-temporal model explaining examination performance over five years in all the 47 counties of Kenya. Previous studies have overlooked space and time attributes in the analyses and this is the motivation for this project. To policymakers, the study will give an understanding of these two critical covariates and explain their influence in student performance across time and space.

## 1.6 Organization of this Project

This project involves model development, computation, and inferences based on a Bayesian spatio-temporal analysis of examination scores data. The two methods are combined to come up with a model that can be used as both a predictive and an analytic tool. Chapter 1 makes an introduction into spatio-temporal methods and Bayesian inference, chapter 2 entails a literature review of the topic under study, chapter 3 gives the background and applications of Bayesian inference and spatio-temporal methods, chapter 4 gives the methodology, and chapter 5 gives the results, discussion, and recommendations for future research.

# CHAPTER 2: LITERATURE REVIEW

## 2.1 Spatio-temporal Modeling

There is a growing body of spatio-temporal models due to the vital need for them due in part to the availability of many datasets having spatial and temporal measures. Spatial measures include altitude, longitude, and distance above sea level (elevation). The increasing availability of applications in the vast field of social science has also necessitated the uptake of spatio-temporal methods. Typical examples of spatial data include rainfall data, epidemiological data and pollution data. In spatio-temporal data, observations are linked to the aforementioned spatial measures and assigned to a specific time during which the data was collected, such as day, month, year, or decades. Applications of spatio-temporal methods include disease mapping, environmental pollution monitoring, weather prediction and control of disease spread. Usually, the interest is to explore spatial and temporal dependence or to come up with predictive models that incorporate the two elements.

The field of spatial statistics has experienced significant growth over the past two decades and this has partly been pushed by the increasing availability and need for spatial data.

Spatio-temporal models are regularly implemented by combining time series models with spatial-based methods. The latter refers to methods used to analyze how data are correlated with respect to time and is used to examine spatial continuity or roughness in data. In a time series context, several approaches have been developed and these include autoregressive moving-average model (ARMA), autoregressive (AR) processes for stationary data and random walk (RW) models. Combined together, space-time methods provide dynamic models that can be used in making more precise forecasts than a purely spatial or temporal model.

## 2.2 School Performance Modeling

Due to the increasing significance of school performance among various stakeholders in the education sector, models of student performance have gained ground in the past few years. These studies have employed logistic models, probit analysis, linear regression, and multilevel logistic regression analysis. The studies have mainly focused on factors such as gender difference, teaching styles employed by teachers, class environment, socio economic factors, and family education background.

Studies into spatial characteristics of student performance have taken different approaches. For instance, Maliki *et al* (2009) examine whether examination performance between students learning in rural schools and urban schools differ. Additionally, they analyze other covariates such as sex, school type (public/private) and performance in mathematics. Using a sample size of 600 students from the Bayelsa state of Nigeria, the authors employed t-test analyses to examine mean differences in performance for different categories. From their findings, the authors conclude that performance differs according to school location. Saha (2011) has also examined the influence of location of school on school examination results. Using a logistic regression approach and subsequent Wald statistic to determine significant coefficients, the authors concluded that performance indeed varies by school location.

Xu (2009) also investigated how various factors affect student performance. Using a sample size of 633 students drawn from both urban and rural schools, Xu embarks on a study aimed at investigating the linkage between student achievement, school location, and homework management strategies. The author contends that students' capacity to regulate their own learning exhibits some spatial pattern in which students attending rural schools are seen to have poor homework management strategies. This is attributed to low educational aspirations among rural students. Additionally, these pupils place less value on academics as compared to their urban counterparts. To explore this concept further, the author collected information among sampled students regarding their homework management strategies. In the exploratory analysis, crosstabs and correlations were used. Exploratory analyses are then used and these include a one-way analysis of variance (ANOVA) which is used to examine mean differences within subjects and a multivariate analysis of variance (MANOVA) to estimate the effects of school location and student achievement on the various categories of homework management.

From this study, Xu finds out that compared to rural school students, urban students have more motivation and engagement during homework. The findings of this study can be used as a generalization to explain poor performance among rural school students. This generalization is based on a study that was conducted by Zimmerman and Kitsantas (2005) in which they investigated the role of students' homework practice in the achievement of specific learning processes, perceptions of academic responsibility, and academic attainment. From the study, the researchers concluded that student academic achievement is positively associated with the quality of homework practices. However, both studies do not include temporal dependence in pupil's performance over time.

Fotheringham, Charlton and Brunsdon (2001) have also investigated spatial variations in school performance using a geospatial analysis method known as geographically weighted regression (GWR). This is a geostatistical procedure put forward by the same authors to model spatial nonstationarity. GWR extends the traditional linear regression structure by allowing local variations in rates of change so that the different model coefficients are used with respect to spatial variations (Brunsdon, Fotheringham and Charlton, 1996).

In application of the GWR, Fotheringham *et al* (2001) examine spatial variations in school performance among 3687 schools in northern England. From the analysis, the authors conclude that there exists a great deal of spatial variation in school performance that cannot be explained by the classic regression framework. In the latter method, the coefficients are only averages across a geographical area and can hide many interesting relationships. Spatial results can help researchers get the actual extent of spatial stationarity or non-stationarity. Additionally, they concluded that the results can help us examine the nature of relationships with respect to other covariates and this would not be possible using a global framework.

Research into factors affecting student performance has also focused on the school catchment area. This is as a realization of the fact that the catchment area is a spatial attribute and is linked to other factors such as socio-economic status. The correlation could be crucial in explaining area profile and school performance (Martin and Atkinson, 2001). Gibbons (2002) used a panel data on primary schools in England to explore this correlation through estimation of the relationship of location, local interactions, and community characteristics to primary school performance. Using a dataset containing basic school characteristics and test scores and another dataset on local area characteristics, the author uses various spatial methods including the K-nearest neighbor to investigate spatial dependence and correlation between school performance and underlying catchment area characteristics. From the study, it is found out that the background of the pupils' influences school performance. It is observed that the distribution of various socio-economic determinants such as parental income and unemployment levels can be linked to school performance. The analysis suffers from one flaw because schools, especially those in which children perform well, usually attract pupils from other neighborhoods or locations. For such schools, pupil performance may not be directly linked to the characteristics of their neighborhoods unless the catchment area is sufficiently comprehended.

Aside from spatial analysis, the inclusion of time-series methods in the analysis of KCPE exam scores has also been important in explaining the observations over the years. For instance, from 2003, the number of pupils sitting for the examinations increased sharply following the introduction of free primary education in Kenya. For instance, standard 1 intake increased from 970,000 in 2002 to 1,300,000 in 2003, resulting in a 35% increase. Prior to the introduction of FPE, intake had been constant for close to 10 years. Enrolment among persons sitting for KCPE examination also increased although not by the same margin as that observed in class 1.

Further temporal analysis of KCPE results has also shown a drop in performance among candidates who sat for the examination in 2007. To investigate this observation, Ogeto (2012) studied the impact of post-election violence (PEV) on pupils' performance in KCPE in public primary schools in Esise division, Nyamira county, Kenya. Using a sample size of 360 pupils, the author concludes that PEV had negative impact on learners' enrolment which decreased gradually from 2535 in 2007 to 2179 in the year 2012. The PEV also had an impact of teaching staff. Generally, the effect of the PEV was a progressive decline in performance from 2008, 2009 and 2010.

## 2.3 Bayesian Methods

The use of Bayesian methods in modeling of spatial data gained attention in the 1990s due to the availability of simulation methods that were incorporated into software such as R and WinBugs. Previously, analysis of spatial data had been based on spatial modifications to the linear regression model. Under this method, for instance, spatial modeling was done though the response variable and a number of covariates. The advent of MCMC methods in Bayesian computation resulted into other simpler yet effective ways of modeling spatial dependence. Recent developments have further enhanced spatial modeling. These developments include that of the Integrated Nested Laplace Approximation (INLA), which has been developed by Rue (2007) to model data drawn from both a Gaussian and non-Gaussian field.

Bayesian methods have gained application in various disciplines among them the analysis of examination results. Moussavi and McGinn (2009) describe a Bayesian Network model to diagnose the causes of low effectiveness of certain schools. Generally, studies on student performance have revealed that school performance differs from region to region while at the individual student level, it differs based on gender and other demographic, socio-economic, and cultural factors.

The aim of this study is to build a tool that can be used by policymakers for two purposes:

i. to explain learning outcomes in terms of conditions and latent processes within schools; and
ii. to estimate the probabilities that given interventions will affect those conditions and processes and their influence on learning outcomes.

# CHAPTER 3: METHODOLOGY

## 3.1 Background on Gaussian Models

Spatio-temporal models are part of a group collectively known as Latent (unobservable) Gaussian models (LGMs). In this project, latent variables include spatial aspects of the dataset, i.e. the spatial element attached to the response variable (county). LGMs are a flexible and widely used class of statistical models. The fundamental part of these models is an unobserved multivariate Gaussian random variable x, whose density $\prod(\mathrm{x}|\theta)$ is determined by a vector of parameters $\boldsymbol{\theta}$. Some of the elements in the random vector $x$ are observed indirectly from data $\boldsymbol{y}$. An assumption of conditional independence is made on data $\boldsymbol{y}$ with respect to the latent field $\boldsymbol{x}$**,** i.e.

$$f(\mathrm{y}|\mathrm{x}) = \prod f(y_i|x_i) \tag{3.1}$$

The elements of a latent Gaussian model are then taken as:

- the likelihood of the data $f(\mathrm{y}|\mathrm{x})$;
- the Gaussian density of the random vector $\boldsymbol{x}$**,** $f(\mathrm{x}|\theta)$;
- the prior distribution of the parameter vector $f(\theta)$;

the posterior distribution can then be obtained as follows:

$$f(\mathrm{x}, \theta|\mathrm{y}) \propto f(\theta)\, f(\mathrm{x}|\theta) \prod f(y_i|x_i) \tag{3.2}$$

It is assumed throughout the computation process that the parameters of interest are found in the posterior marginal for $\boldsymbol{x_i}$**.**

The latent Gaussian field $\boldsymbol{x}$ presents a flexible process that can be used to include spatial and time dependence among other potential covariates in a statistical model. Previously, the typical tool for making Bayesian inference on Gaussian field-based models was the Markov Chain Monte Carlo (MCMC). However, the complexity of the latent field, the (frequently) high dimensionality of the latent field $x$, and the strong correlation within $\boldsymbol{x}$ and between $\boldsymbol{x}$ and $\boldsymbol{\theta}$ results into problems during the convergence process and in the mixing properties of the Markov chain. Although block update strategies have been devised as a solution to this problem, MCMC methods still remain significantly slow. A new method based on the deterministic approximations to the posterior marginals of interest has been devised that completely evade the MCMC process. This new approach offers several advantages over the conventional MCMC approach and the main one is the computational advantage- since the approach is simulation free, the parameters of interests can be obtained in a few seconds or minutes while for a similar computation, MCMC algorithms would require hours or even days. In addition, the new approach is more accurate with respect to the computational time as compared to MCMC processes that would have to run for a much longer time for any bias to be identified. This method is referred to as integrated nested Laplace approximation (INLA).

INLA was proposed by Rue, Martino and Chopin (2009) to perform approximate fully Bayesian inference on the class of latent Gaussian models (LGMs). It makes use of deterministic nested Laplace approximations and, as an algorithm tailored to the class of LGMs, it provides a faster and more accurate alternative to simulation-based MCMC schemes.

Under INLA, the approximation of posterior marginals of the latent Gaussian field, $\mathrm{f}(x_i|\mathrm{y}),\ \mathrm{i} = 1, \ldots, \mathrm{n}.$ proceeds in three steps:

- The initial step approximates the posterior marginal of θ using the Laplace approximation;
- In the second step, the simplified Laplace approximation, $\prod(x_i|\mathrm{y}, \theta)$ is computed for particular values of θ;
- At the final step, the first two steps are combined through numerical integration;

The usage of MCMC methods for dynamic computations such as spatio-temporal or time-series models is only possible through simulation and requires complicated sampling schemes to ensure efficiency (Gamerman, 1998). Furthermore, the inclusion of random effects such as spatial covariates might require an extensive reparametrization of the model. On the other hand, INLA can be run within a user-friendly R environment with generic functions. All code is available freely available from www.r-inla.org and from http://cran.r-project.org/ (R Development Core Team). All analyses in this project were done using INLA version 0.0 on R-package version 3.0.1.

### 3.2 Bayesian Inference

The foundation of Bayesian statistics is Bayes' theorem. Suppose we observe a random variable $y$ and wish to make inferences about another random variable $\theta$, where $\theta$ is drawn from some distribution $p(\theta)$. From the definition of conditional probability,

$$p(\theta|y) = \frac{p(y \cap \theta)}{p(y)} \tag{3.3}$$

From the definition of conditional probability, we can express the joint probability as follows:

$$p(y \cap \theta) = p(y|\theta)\, p(\theta) \tag{3.4}$$

Combining these two formulas gives Bayes' theorem

$$p(\theta|y) = \frac{P(y|\theta)\, p(\theta)}{p(y)} \tag{3.5}$$

The formula contains three important elements:

$p(\theta)$ : this is the prior distribution of the possible $\theta$ values and is the probability that a model is true before any data are observed. It must come from prior information, not from the current data

$p(\theta|y)$ : this is the posterior distribution of $\theta$ based on the observed data or current information and this refers to the probability that a model is true after observed data have been taken into account.

$p(y)$ : this is the marginal distribution of the data or the normalizing constant; and

$p(y|\theta)$ is the sampling density for the data—which is proportional to the Likelihood function, only differing by a constant that makes it a proper density function (normalizing constant);

The *prior* provides information that exists prior to the estimation process and its inclusion ultimately produces a posterior probability that is also no longer a single quantity; instead, the posterior becomes a probability distribution as well.

Bayes' theorem can also be expressed in terms of probability distributions as follows:

$$f(\theta|y) = \frac{f(y|\theta)f(\theta)}{f(y)} \tag{3.6}$$

where $y$ represents the data.

For a continuous process, the marginal distribution, sometimes referred to as the marginal likelihood, or simply the likelihood, can be written as:

$$f(y) = \int f(y|\theta)f(\theta)d\theta \tag{3.7}$$

Computation of the likelihood frequently presents problems as it often involves very large summations or multidimensional integral and this inhibited the development of Bayesian methods for almost half a century. Consequently, computations were limited to simple problems in which the integration was tractable. However, in the early 1990s, a solution was found in form of the Markov Chain Monte Carlo process (MCMC) in which integration was conducted through simulation. For instance, if we are to create a sample form the posterior distribution, then the integral:

$$g(x) = \int m(\theta)\, f(\theta|\mathrm{y})d\theta \tag{3.8}$$

can be approximated by:

$$\hat{g}(x) = \frac{\sum_{i=1}^{n} m(\theta_i)f(y|\theta_i)f(\theta_i)}{\sum_{i=1}^{n} f(y|\theta_i)f(\theta_i)} \tag{3.9}$$

where the function $m = m(\theta)$ provides the posterior parameter that the researcher is interested in, such as the posterior mean.

The marginal is used as a normalizing constant to make the posterior density proper. Hence, the posterior, likelihood and the prior are normally linked proportionally as shown:

$$Posterior\ \alpha\ Likelihood\ x\ Prior \tag{3.10}$$

The following is a general process for fitting Bayesian models:

a) A probability distribution for $\theta$ is formulated as $f(\theta)$, which is known as the *prior*. The prior distribution expresses the researcher's beliefs about the parameter of interest before examining the data;

b) Given the observed data y, the researcher choose a statistical model $p(y|\theta)$ to describe the distribution of $y$ given $\theta$;

c) The researcher then computes the posterior distribution and hence updates his belief about $\theta$ by combining information from the prior distribution and the data $p(\theta|y)$. This is used to describe the conditional probability of the data given a particular model.

### 3.2.1 Prior Distributions in Bayesian Methods

A prior distribution of a parameter is the probability distribution that represents a researcher's uncertainty about the parameter before the current data are examined. Multiplying the prior distribution and the likelihood function together leads to the posterior distribution of the parameter. We can use the posterior distribution to carry out all inferences but cannot carry out any Bayesian inferences or perform any modeling without using a prior distribution.

The choice of a prior has always been a source of controversy, with frequentists asserting that the insistence on a prior could somehow result into biasness in the computation process, or that the choice of a prior depends on the researcher's intentions and that the benefits of Bayesian computation are negated by the requirement to state a prior. Fortunately, the fears are not based on facts. Priors are simply an expression of the researcher's degree of belief in the parameters in the absence of new data. Priors are critical for continuous scientific knowledge and for improving results from small sample studies. As a field of study matures, more and more data is made available and this influences the type of priors chosen in successive experiments, resulting into more precise and well-founded conclusions. It is unreasonable to omit priors as they have little influence on the posterior, especially when little information is available regarding the parameters. The resulting posterior is normally a tradeoff between the prior and the collected data.

Different priors can be adopted depending on current knowledge regarding the parameters and on the likelihood. For example, consider an ongoing vaccination process. In this case, choice of prior to model the disease outbreak will be well informed and will heavily influence the posterior. However, in an initial vaccination, information on priors will be vague and thus their influence on the prior will be limited. Indeed, the choice of a prior must be based on the evidence available, either through previous research, literature reviews, or the researcher's own judgment of the situation.

Cox and Hinkley (1974) have proposed three techniques for coming up with priors:

a) An empirical Bayesian approach proposes that the prior should be based on previous data
b) An objective approach proposes the prior should be based on rational belief about the parameter, or on mathematical properties, that in some sense maximizes information gain (Berger, 2006).
c) A subjective approach proposes that the choice should simply quantify what is known or believed before the experiment takes place, and is an expression of the level of belief by the researcher;

Various types of priors can be used to cater for different situations regarding the researcher's level of belief. These include objective/subjective priors, informative priors, non-informative priors, improper priors, conjugate priors, and Jeffrey's priors. A Gaussian prior that can be regarded as an informative prior will be used in this project.

### 3.2.2 Challenges in Bayesian Inference

A historical problem that has previously hindered adoption of Bayesian methods is the computation of the marginal distribution, $p(y)$. One solution has always been to use conjugate priors, i.e. priors that make the posterior function come out with the same function as the prior. The computation process is not complicated. Examples of conjugate priors/likelihood are:

| **Prior** | **Likelihood** |
|---|---|
| Normal | Normal |
| Gamma | Poisson |
| Gamma | Gamma |
| Gamma | Beta |
| Beta | Binomial |
| Dirichlet | Multinomial |
| Normal ($\mu$ unknown, $\sigma^2$ known) | Normal |

| | |
|---|---|
| Normal (μ unknown, $\sigma^2$ unknown) | Normal |
| Multivariate normal (μ unknown, V known) | Multivariate normal |

A second solution is to approximate the actual functions with other functions that are easier to work with, and then show that the approximation is reasonably good under typical conditions.

A third problem normally encountered is in the determination of a practical prior. For a start, a researcher should use a plain or vague prior and update the parameters as more information becomes available. In the event that scientists cannot agree on the prior to use, then different priors can be used and then the robustness of the posterior can be assessed against changes in the prior. Another solution is to mix the two priors to create a joint prior that accounts for uncertainty. In summary, for most applications, specification of the prior turns out to be technically unproblematic, although it is conceptually very important to understand the consequences of one's assumptions about the prior. Thus, the main reason that Bayesian analysis can be difficult is the computation of the likelihood. However, the computation is tractable in many situations via a number of methods such as Markov chain Monte Carlo (MCMC) and through the integrated nested Laplace approximation (INLA).

The use of Bayesian methods in this project is to take represent prior uncertainties about model parameters with a probability distribution and to update this prior information with current data that produces a posterior distribution for the parameter that contains less uncertainty. Consequently, a prior distribution and a likelihood distribution will be assigned and using the data on student scores, the prior information regarding these parameters will be updated.

### 3.3 Data description

The data used for this study refers to the KCPE scores of all primary schools in the 47 counties of Kenya from 2006 to 2010. Additionally, the data contains KCPE scores from South Sudan. The scores have been collapsed into different categories such as gender, range of marks, county, and year. The number of pupils in each category is also given.

In view of the scope of this project, a response variable was created out of the row data. The aim was to conduct a spatio-temporal analysis of the number of students attaining 350 marks and above. This cutoff mark used was because in most counties in Kenya, a score of 350 marks (out of 500) guarantees admission into a good secondary school. The response variable was created by aggregating the number of students scoring 350 marks and above for every county in each of the five years. This resulted into lattice data pertaining to all 47 counties in Kenya. Data on the performance of students from South Sudan was dropped during the data management process.

## 3.4 Models used in the Project

This project concentrates on spatio-temporal modeling and associated inference with the aim of establishing a framework, where student performance in different parts of the country can be quantified through space and time. The model involves a Gaussian Field (GF), affected by a measurement error, and a state process characterized by a first order autoregressive dynamic model and spatially correlated covariates. Traditionally, this kind of relation has been efficiently modeled through Markov chain Monte Carlo (MCMC) techniques within the Bayesian framework, which pose computational problems and also take hours to converge. The goal of the project is to put forward an efficient estimation and spatio-temporal prediction approach.

This application consists in representing a GF with Matern covariance function as a Gaussian Markov Random Field (GMRF). The Matern covariance function is used to define statistical covariance between two points that are *d* units from each other. The covariance function is in itself a function of the distance between a geographical point or area and its neighbors and it is used to link elements within a random field. For this project, this is represented by the distance between the central point in a county and that of its neighbors. A big advantage of moving from a GF to a GMRF stems from the good computational properties that the latter enjoys. Additionally, when dealing with Bayesian inference for GMRFs, it is possible to employ the Integrated Nested Laplace Approximation (INLA) approach as opposed to the computationally bulky MCMC methods resulting into computational advantages (Cameletti, Lindgren, Simpson & Rue, 2011b).

This methodology employs the following models: *bym* model, *besag* model, and the *iid* model. These models are explained below:

### 3.4.1 *Besag* model for spatial effects

The proper version of the *besag* model for random vector $\boldsymbol{x} = (x_i, \dots x_n)$ is defined as

$$x_i | x_{\mathrm{j}}, i \neq j, \tau \sim N\left(\frac{1}{n_i}\sum_{i\sim j} x_j, \frac{1}{\tau n_i}\right) \tag{3.11}$$

where $n_i$ is the number of neighbours of node $i$, $i\sim j$ indicates that the two nodes $i$ and $j$ are neighbours, $d > 0$ is an extra term added on the diagonal controlling the “properness" $\tau > 0$ is a “precision-like" (or scaling) parameter.

### 3.4.2 Model for correlated random effects: *iid*

The *iid* model is used to represent the correlated random effects arising from temporal effects. The model accounts for heterogeneity across time, i.e. the random effect of time on the observations. The random effects model is more efficient in spatial statistics than a fixed effects model because of correlations among observations and due to the random effects of time. This randomness arises from the contribution of several temporal attributes on the observations. In the INLA package, the model is specified as follows:

*y ~ f(i, model="iid",n = <length>) + ...*

### 3.4.3 Random walk of order 1: *rw1*

This study will use one type of latent model known as random walk which is a random process consisting of a sequence of discrete steps of fixed length. A distribution is said to follow a random walk if the first differences (difference between two successive observations) are random. In a random walk model, the series itself is not random, however, its differences are. The differences are independent, identically distributed random variables with a common distribution. The implication of using this model is that examination score in any one year in a county depends largely on the results of the previous year.

A random walk of order 1 model is defined as follows:

$$X_t = X_{t-1} + \Delta x_i \tag{3.12}$$

where

$X_t$ is the value in time period t,

$X_{t-1}$ is the value in time period t-1 (one time period before)

$\Delta x_i$ is the value of the error term in time period t.

Since the random walk was defined in terms of first differences, it may be easier to see the model expressed as:

$$X_t - X_{t-1} = \Delta x_i \tag{3.13}$$

In INLA, the random walk of order 1 (RW1) for the Gaussian vector $x = (x_1, x_2, \dots, x_n)$ is constructed assuming independent increments:

$$\Delta x_i = x_i - x_{i+1} \sim N(0, \tau^{-1}) \tag{3.14}$$

In INLA, the *rw1* model is specified inside the f() function as follows:

```
f(<whatever>, model="rw1", values=<values>, cyclic=<TRUE|FALSE>, hyper = <hyper>)
```

### 3.4.4 The Besag-York-Mollie model for spatial effects: *bym*

This model is simply the sum of a *besag* model and an *iid* model. The benefit is that this allows us to get the posterior marginals of the sum of the spatial and *iid* model; otherwise it offers no advantages.

### 3.4.5 Gaussian Prior

Since the spatio-temporal process follows a Gaussian distribution, a similar distribution is attached to the prior.

The normal/Gaussian distribution has density given by:

$$\mathrm{f}(\theta) = \left(\frac{\tau}{2\pi}\right)^{\frac{1}{2}} \exp\left(-\frac{\tau}{2}(\theta - \mu)^2\right) \tag{3.15}$$

for continuous $\theta$ where

$\mu$: is the mean

$\tau$: is precision (thus variance is given by $\frac{1}{\tau}$)

### 3.4.6 Negative Binomial Likelihood

The dependent variable in this analysis, Y, is a count variable and represents the number of students scoring at least 350 marks for each of the 47 counties from 2006 to 2010. The assumption we make is that Y follows a negative binomial distribution, and this is the likelihood. Since Y is a count variable, a Poisson distribution could be used too. However, preliminary analysis of the data shows that it does not meet the requirements for a Poisson process, i.e. mean and variance are not equal. The response variable is assumed to be independently and identically distributed for all areas, i.e. does not follow any spatial patterns (unstructured component). A second assumption for the model is that it belongs to the Gaussian family. This assumption is validated by the principle that every distribution can be represented as a Gaussian distribution.

The negative Binomial distribution is

$$f(y) = \frac{\Gamma(y+n)}{\Gamma(n)\,\Gamma(\mathrm{y}+1)} p^{n}(1-p)^{y} \qquad (3.16)$$

for responses y = 0,1,2,…, where

$n$: number of successful trials (size), or dispersion parameter. This must be positive but not necessarily an integer. This refers to the number of students scoring ≥350 marks in every county

$p$: probability of success in each trial- refers to the probability of scoring ≥350 marks in any county;

In INLA, the negative binomial likelihood is specified as follows:

*family = nbinomial*

Required arguments: *y* and *E* (by default *E* = 1)

The two arguments *y* and *E* represent the response variable and expected values (for each county) respectively.

## 3.5 The Spatio-temporal Model

The model consists of both time and space elements. Time is assumed to follow a random walk of order 1 (RW1) process while space will be modeled using the aforementioned *bym* model.

The general model accounting for both the spatio and temporal attributes can thus be stated as follows;

$$Y(s_i, t) = z(s_i, t)\beta + \varphi(s_i, t) + \epsilon(s_i\,,\, t) \tag{3.17}$$

where

$\varphi(s_i, t)$ is the temporal effect

$z(s_i, t)\beta$ is a process based on fixed effects, i.e. independent of both space and time

$\epsilon(s_i\,,\, t)$ is an error term

In INLA, the model incorporating the *iid*, *rw1*, and the *besag* models is formulated as;

$$y{\sim}1 + f(time,\ model = rw1) + f(time,\ model = iid) + f(area,\ model = besag)$$

Since the *bym* model is a sum of the *besag* and *iid* modes, the models can also be formulated as,

$$y{\sim}1 + f(time,\ model = rw1) + f(time,\ model = iid) + f(area,\ model = besag)$$

In the model, $\beta(s_i, t) = 1$ since no covariates were included in the analysis while $\alpha(s_i, t)$ is represented by the *bym* model. The *bym* model is a sum of a *besag* model and an *iid* model that have been explained in earlier sections. These two models account for spatial and time dependence respectively. Additionally, a second model is used in which time is assumed to follow a *rw1* process in which differences between subsequent observations are assumed independent, identically distributed random variables with a common distribution.

## CHAPTER 4: DATA ANALYSIS AND DISCUSSION

### 4.1 Exploratory Data Analysis

We begin by justifying the inclusion of time and space in our model. In this analysis, we explore the significance of both time and space covariates through a time-series plot of the response variable and a scatterplot of Y against the size of each county respectively.

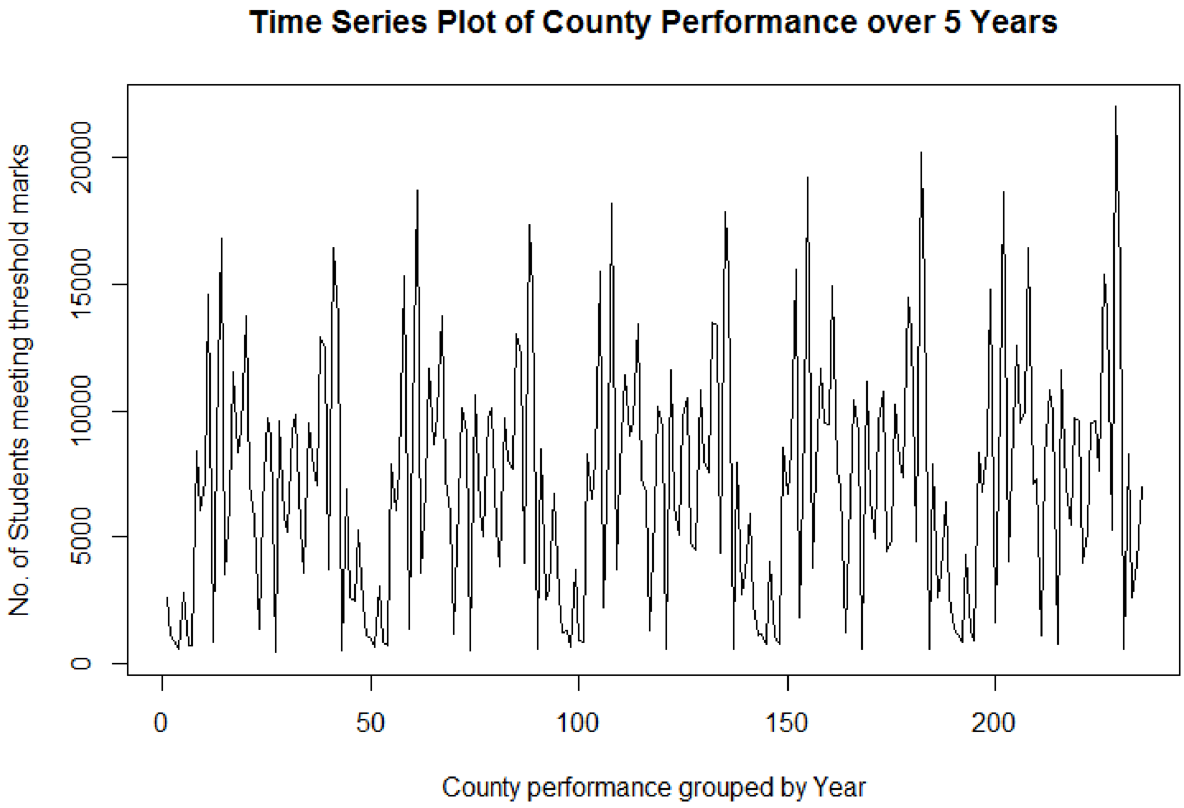

Fig 3. Time series plot of students obtaining threshold marks in 47 counties over five years

Next, we justify the inclusion of location attribute in the model. This is done by drawing a two-way scatterplot of the dependent variable against area of each county.

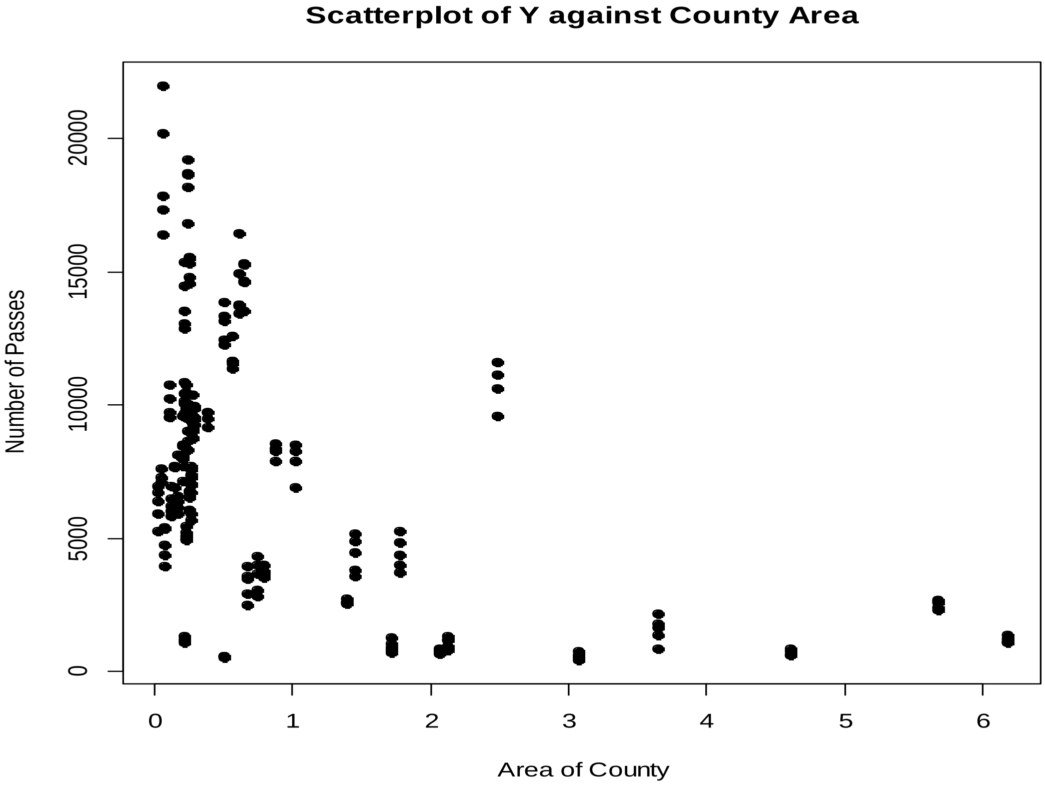

Fig. 4. Scatterplot of students obtaining threshold marks per county against area of county

The correlation between Y and Area of County was also calculated and found as -0.505. A histogram for the response variable is shown below.

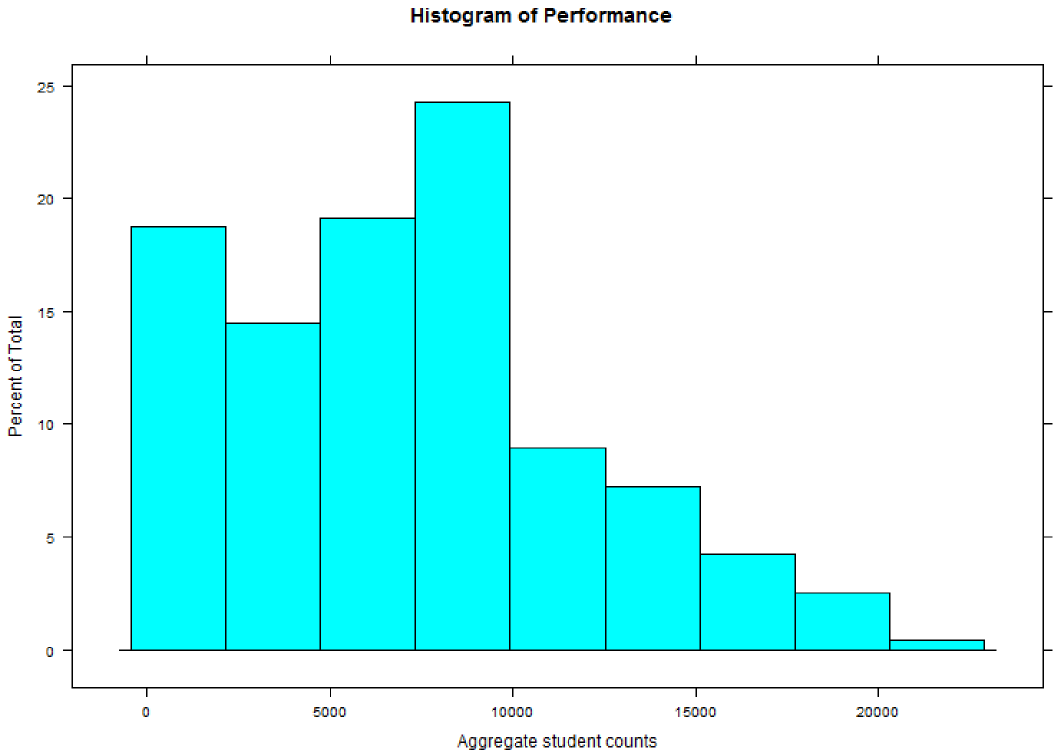

Fig. 5: Distribution of students obtaining threshold marks per county

The final section of this exploratory analysis involves a summary statistic for the dependent variable and the results are as shown below:

| observations | mean | median | s.d. | minimum | maximum |
|---|---|---|---|---|---|
| 235 | 7200.106 | 7103 | 4788.44 | 435 | 22013 |

## 4.2 Confirmatory Analysis

This section focuses on space-time modeling using the various models stated earlier. The first model assumes that the time process follows an *rw1* process while the second model makes an added assumption that the time process follows an *iid* process. A Gaussian prior with parameters (0,0.01) and negative binomial likelihood distributions are used. The use of a negative binomial for the likelihood is because the response variable, Y, refers to count data, i.e.  number of students scoring at least 350 marks per county. For count data, Poisson likelihood could have been used but the requirement for such a model is that the mean must be equal to the variance. This condition is not met since the mean for Y is 7200 while the variance is 22929148. Consequently, negative binomial distribution likelihood is assigned to Y.

Four models were fitted to the data and subsequently the deviance information criterion (DIC) was used to determine the best fit. The models are shown below

Model 1: bym model + rw1 model

```
Fixed effects:

             mean     sd 0.025quant 0.5quant 0.975quant kld
(Intercept) 0.0843 0.0404     0.0064   0.0838      0.165   0

Random effects:
Name          Model
ID.area   BYM model
year   RW1 model

Model hyperparameters:
                                                    mean      sd        0.025quant 0.5quant  0.975quant
size for the nbinomial observations (overdispersion) 2.679e+00 2.438e-01 2.167e+00  2.697e+00 3.106e+00
Precision for ID.area (iid component)               1.766e+03 1.802e+03 1.052e+02  1.224e+03 6.512e+03
Precision for ID.area (spatial component)           2.024e+02 1.787e+03 4.875e+00  3.362e+01 1.304e+03
Precision for year                                  1.890e+04 1.863e+04 1.347e+03  1.343e+04 6.830e+04

Expected number of effective parameters (std. dev): 14.33(6.723)
Number of equivalent replicates : 16.40

Deviance Information Criterion: 4471.34
```

## Model 2: bym model + rw1 model + iid model

```
Fixed effects:
            mean     sd 0.025quant 0.5quant 0.975quant kld
(Intercept) 0.0828 0.0408     0.0044   0.0822     0.1645   0

Random effects:
Name           Model
ID.area   BYM model
ID.year   RW1 model
ID.year1   IID model

Model hyperparameters:
                                                    mean      sd        0.025quant 0.5quant  0.975quant
size for the nbinomial observations (overdispersion) 2.831e+00 2.754e-01 2.308e+00  2.826e+00 3.387e+00
Precision for ID.area (iid component)              2.439e+03 2.903e+03 2.056e+02  1.565e+03 9.997e+03
Precision for ID.area (spatial component)          4.374e+01 1.295e+02 3.104e+00  1.600e+01 2.545e+02
Precision for ID.year                              2.142e+04 2.367e+04 1.774e+03  1.433e+04 8.366e+04
Precision for ID.year1                             2.485e+04 2.964e+04 2.015e+03  1.591e+04 1.022e+05

Expected number of effective parameters (std. dev): 15.69(7.675)
Number of equivalent replicates : 14.98

Deviance Information Criterion: 4471.34
```

## Model 3: rw1 model + iid model

```
Fixed effects:
            mean     sd 0.025quant 0.5quant 0.975quant kld
(Intercept) 0.094 0.0403     0.0157   0.0937     0.1739   0

Random effects:
Name          Model
ID.year   RW1 model
ID.year1   IID model

Model hyperparameters:
                                                   mean      sd        0.025quant 0.5quant  0.975quant
size for the nbinomial observations (overdispersion) 2.685e+00 2.381e-01 2.256e+00  2.671e+00 3.191e+00
Precision for ID.year                              1.902e+04 1.866e+04 1.345e+03  1.354e+04 6.836e+04
Precision for ID.year1                             1.870e+04 1.856e+04 1.212e+03  1.320e+04 6.747e+04

Expected number of effective parameters (std. dev): 1.114(0.0933)
Number of equivalent replicates : 210.98

Deviance Information Criterion: 4479.89
```

## Model 4: besag model

```
Fixed effects:
             mean     sd 0.025quant 0.5quant 0.975quant kld
(Intercept) 0.0828 0.0401     0.0058   0.0821     0.1634   0
```

```
Random effects:
Name          Model
ID.area   Besags ICAR model

Model hyperparameters:
                                              mean     sd       0.025quant 0.5quant 0.975quant
size for the nbinomial observations (overdispersion)   2.8987   0.2982   2.3184     2.9010   3.4840
Precision for ID.area                         63.5997 291.1838   3.0347    17.1225 398.0938

Expected number of effective parameters (std. dev): 15.60(7.916)
Number of equivalent replicates : 15.07

Deviance Information Criterion: 4471.52
```

| **Model** | **DIC** |
|---|---|
| bym + rw1 | 4471.34 |
| bym + rw1 + iid | 4471.34 |
| rw1 + iid | 4479.89 |
| besag | 4471.52 |

The posterior means of student scores was computed over the five-year period and for each county. The 2 plots are shown below;

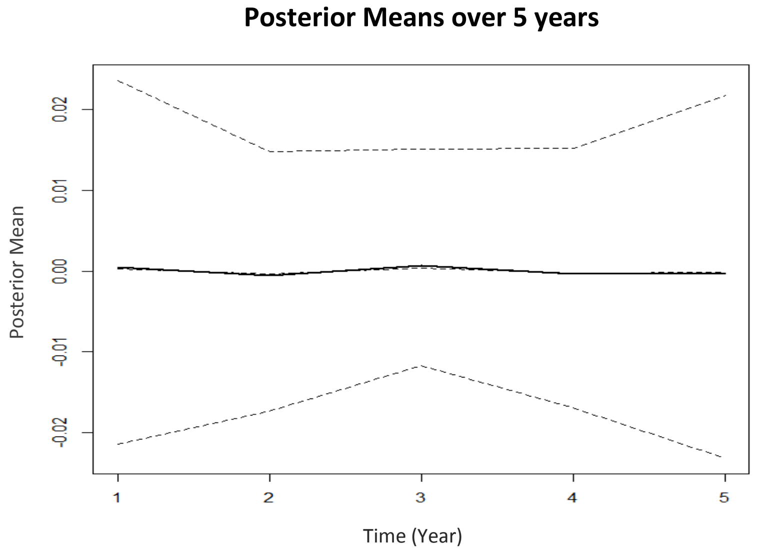

Fig. 6: Means of the posterior distribution over the five-year period

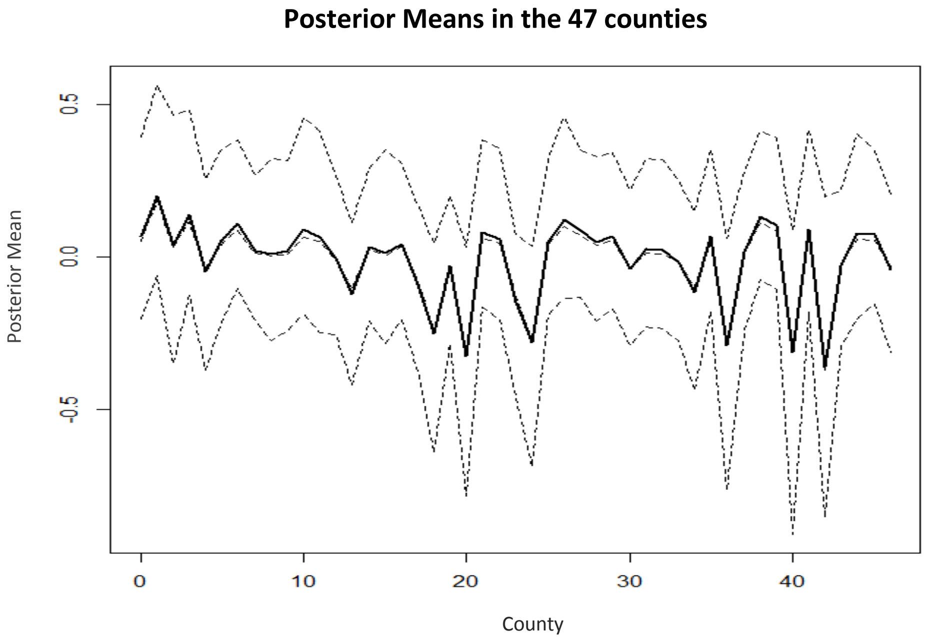

Fig. 7: Means of the posterior distribution in the 47 counties

The posterior means of student scores was computed in the 47 counties and the plot is as shown below;

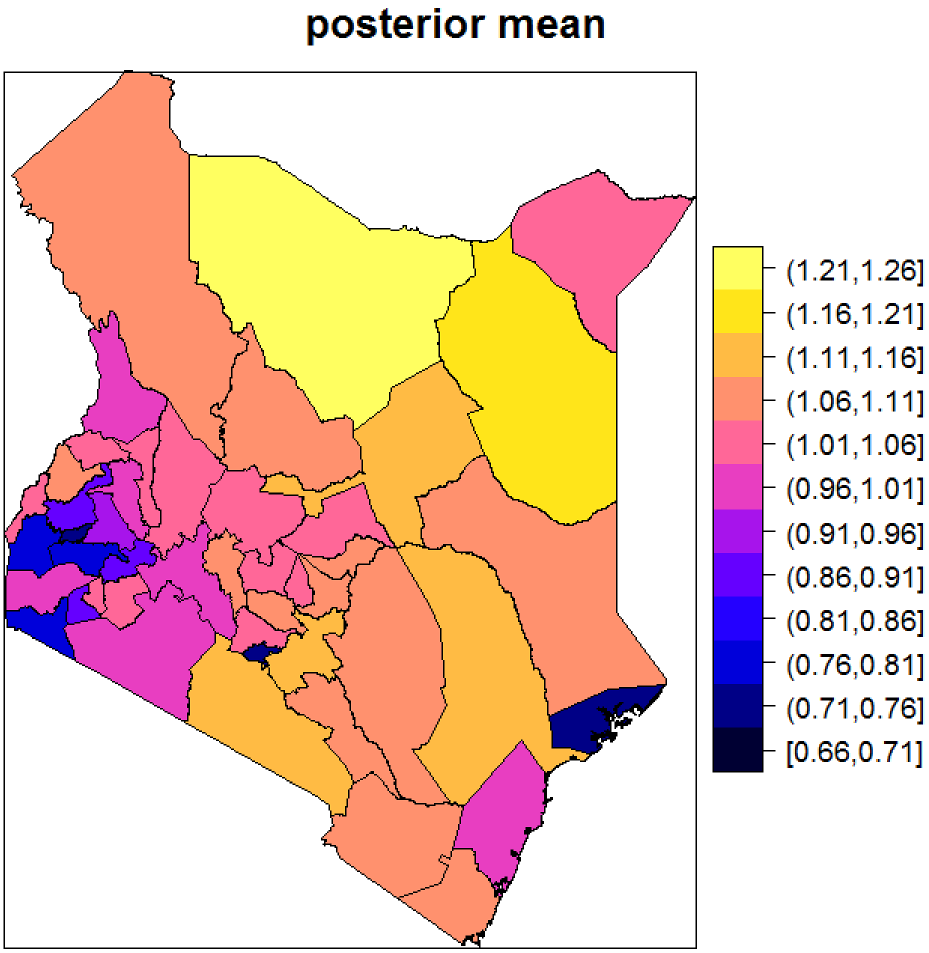

Fig. 8: Map showing standardized posterior means per county

In line with this study, a national trend plot indicating the probability of a student scoring the threshold marks (greater than 350) in each county was constructed. In Bayesian inference, focus is normally on the posterior parameters (posterior mean, or $\theta$, in this case). Consequently, the trend plot below shows the probability that the mean (based on standardized z-scores) is greater than one. This corresponds to the probability of a student scoring the threshold marks in any county and is as shown.

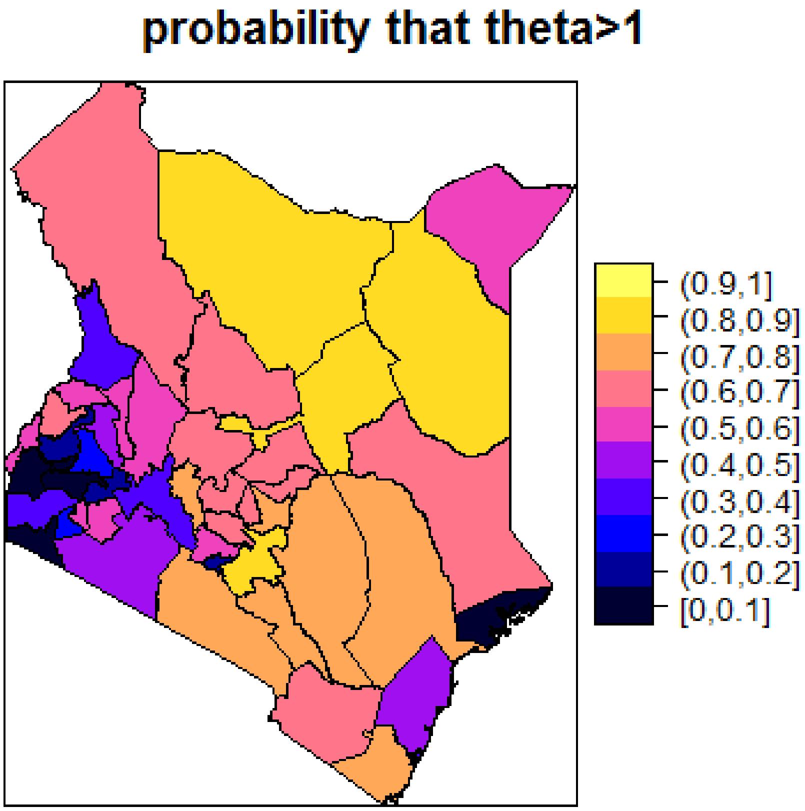

Fig. 9: Map showing probabilities of scoring threshold marks per county

### 4.3 Discussion

Figure 3 shows a time-series model used to justify the use of a time series model. A time series model is justified whenever a temporal ordering is observed in the data, that is, a cyclic trend or regularity in the data at regular time intervals. From the diagram, a regular trend is observed in each of the five years. The pattern is recurring but the magnitude is increasing every year. The increase in magnitude implies an increase in the number of pupils obtaining the threshold marks as defined in the project.

In order to show significance of location/area in analysis, a plot of the response variable against county area was made as shown in Figure 4. The plot shows a negative correlation between the two. Normally, a strong positive correlation is expected since an increase in geographical size implies an increase in the number of student talking examination. However, it is known that in Kenya, population density in any county increases with decrease in size. For instance, from the 2009 population census, Garissa county has a size of 45,720 km sq and a population of 623,060 persons while Kisii county, with a size of 1,317 km sq has a population of 1,152,282 persons. Consequently, there exists a negative correlation between size of county and the number of pupils sitting for examination in Kenya. The existence of a correlation (-0.51) between county size and number of pupils obtaining the threshold marks implies that the incorporation of spatial attributes in the model would increase its predictive and analytic role.

Further exploratory analysis of the data through summary statistics and scatterplot is shown. The summary is made to determine the likelihood distribution. The mean is found to be 7200.16 while the standard deviation is 4788.44. The smallest number of students scoring at least 350 marks was recorded in Kajiado (435) while the highest number was recorded in Embu (22,013). Generally, most counties had observations ranging between 1000 and 11,000 as shown in the plot in Figure 6.

The confirmatory analysis involved computing the deviance information criterion (DIC) to select the best fit for the data. The DIC values are based on a trade-off between model complexity and the fit. Hence, a smaller DIC value implies less trade-off and hence a better fit.

The models are as follows;

Model 1

The model was a sum of the *bym* model and *rw1* model. As stated earlier, the *bym* accounts for both spatial and time dependence (*besag + iid*). The model also included the assumption that followed an *rw1* process.

Model 2

The model was a sum of the *bym*, *iid* and *rw1* model.

Model 3

This model is a sum of the *rw1* and *iid* models. This is a purely temporal model and is aimed at examining whether time alone is enough to explain the dependent variable.

Model 4

This model only includes the *besag* model and the objective of its formulation is to examine whether space alone could explain the dependent variable. As opposed to the second model, this is a purely spatial model.

As expected, the DIC values for model 1 and model 2 were equal since the *bym* model also included the *iid* model. Consequently, both models are a sum of the *bym*, *rw1* and *iid* models. Both models had a DIC value of 4471.34.

Model 3 only accounted for temporal effects and its formulation was meant to examine whether space was insignificant in explaining the number of students scoring at least 350 marks. The DIC value for the temporal model was 4479.89.

Model 4 only accounted for spatial effects and its formulation was meant to examine whether space alone could be used to explain the dependent variable. This model had a DIC value of 4471.52.

The model with the lowest DIC value should be the best fit for the data. Model 1 and model 2 have the lowest values and are hence chosen. However, since model 2 is just a repeat of model 1 albeit with an additional model for the *iid* process, we chose latter model over the former. Hence, the best model for the response variable is a sum of the *bym* and *rw1* model i.e. a sum of both spatial and temporal effects. Further analysis in this study will employ this model.

Model 1 is made up of three components: the random component (*rw1*), the spatial effect component (*besag*), and the time component (*iid*). From the output, the spatial component has the smallest precision component followed by the temporal component and last is the random effect component. Since variance is given by the reciprocal of precision, the spatial component has the largest variance while the temporal component has the least variance. Consequently, the temporal component is most significant in modeling and predicting pupil scores in KCPE examination.

The significance of temporal effects in student performance has been shown in numerous studies. Indeed, most schools attempt to predict exam results or scores based on the results of subsequent year(s). In most cases, result of any year is strongly related to that of the previous year and this is shown in our model. Models explaining examination performance should therefore include temporal attributes

Analysis of the posterior means over the five-year period is shown in Figure 6. The plot shows stability in performance over the period with a slight increase in 2008. Ogeto (2012) has found out in his paper that performance in KCPE dropped in 2008 due to post-election violence (PEV) and this led to a progressively declining negative index in the 2008, 2009 and 2010 examinations. However, it is known that the PEV began more than a month after pupils had cleared their examination, that is, from December 2008 to February 2009. Consequently, the PEV could not have had an effect on examination results. Besides, the sample used by Ogeto was only drawn from Esise division, Nyamira county, and is therefore not representative of the entire county. Ogeto makes a similar finding that performance in KCPE dropped progressively from 2008 onwards up to 2010. Further research would be recommended to investigate the effect of the campaign period and seubsequent PEV on performance in KCPE.

Analysis was also done on the posterior means over the 47 counties and the graph is shown in Figure 7. The graph shows varying levels of performance in each of the counties. For instance, Nakuru (20) and Kajiado (40) counties show a relatively poor performance over the period while Kisumu (25) and Trans Nzoia (10) have comparatively higher number of students attaining the threshold score. The plot also shows mid-sized counties recording higher numbers of students as compared to smaller counties, i.e. from county no 30 onwards. A complete list of the numbering used for the 47 counties is shown in the appendix.

In line with the objective of this project, a plot of the posterior means for the entire country was made. From the map in Figure 8, counties in the central part of the country are seen to have higher means as compared to other parts of the country. These counties include Kirinyaga, Embu, Meru, Nyandarua, and so on. In contrast, counties located on the western part of the country such as Kisumu, Siaya, Homa Bay, and Busia and those located in the coastal region such as Lamu and Kilifi have lower means. Various studies, including a four-year survey by the Millennium Cities Initiatives, have found that infant mortality rates are highest in areas with low academic performance. Specifically, the lake and coastal regions have been observed to have high mortality rates as compared to other parts of the county. Hence, it can be hypothesized that a correlation exists between infant mortality and academic performance.

Further analysis of the probability of a student obtaining at least 350 marks is made as shown in Figure 9. The map is an extension of the previous one in which the posterior means for each county were computed and plotted. Similar to previous findings in this project, it is observed that pupils enrolled in schools around the lake and coastal regions have the lowest probabilities of obtaining at least 350 marks in the KCPE examinations. The highest probabilities are observed in counties around the central part of the county.

## 4.4 Implications for Policy and Future Research

A direct implication of findings from this project is that education should be promoted in the areas that recorded low numbers of students scoring 350 marks and above. These are the lake and coastal regions. Interventions could include construction of more classrooms or schools, hiring of more teachers and encouraging more parents to take their children by offering incentives such as school feeding programs.

This research demonstrated the need to incorporate spatial and temporal attributes when creating models for school performance. A limitation of the research was that analysis was at county level while it is known that examination results are not homogenous in any county. Consequently, useful information was lost in the process of aggregating the data at county level. Hence, in future research, analyses should be based on school-level data as opposed to county-level.

## Appendix

County dimensions and numbering as used in the study

| COUNTY ID | COUNTY | County Length | County Area |
|---|---|---|---|
| 1 | Turkana | 15.04684 | 5.676985 |
| 2 | Marsabit | 11.97416 | 6.176831 |
| 3 | Mandera | 7.355154 | 2.117196 |
| 4 | Wajir | 9.838408 | 4.609589 |
| 5 | West Pokot | 5.030271 | 0.740481 |
| 6 | Samburu | 8.311013 | 1.713014 |
| 7 | Isiolo | 10.18141 | 2.05982 |
| 8 | Baringo | 5.964266 | 0.877177 |
| 9 | Keiyo-Marakwet | 4.063527 | 0.245208 |
| 10 | Trans Nzoia | 2.387265 | 0.201983 |
| 11 | Bungoma | 3.075944 | 0.245741 |
| 12 | Garissa | 11.28113 | 3.641823 |
| 13 | Uasin Gishu | 3.717922 | 0.271894 |
| 14 | Kakamega | 4.08931 | 0.242844 |
| 15 | Laikipia | 5.358738 | 0.786753 |
| 16 | Busia | 2.682712 | 0.147514 |
| 17 | Meru | 4.284992 | 0.56055 |
| 18 | Nandi | 2.986884 | 0.233172 |
| 19 | Siaya | 2.924506 | 0.285189 |
| 20 | Nakuru | 6.474904 | 0.604913 |
| 21 | Vihiga | 1.364576 | 0.045428 |
| 22 | Nyandarua | 3.587821 | 0.265942 |
| 23 | Tharaka | 3.090717 | 0.21696 |
| 24 | Kericho | 2.979811 | 0.170724 |
| 25 | Kisumu | 3.115223 | 0.217291 |
| 26 | Nyeri | 3.050349 | 0.270857 |

| | | | |
|---|---|---|---|
| 27 | Tana River | 11.60308 | 3.074107 |
| 28 | Kitui | 9.028668 | 2.483477 |
| 29 | Kirinyaga | 1.874547 | 0.12001 |
| 30 | Embu | 3.339148 | 0.229668 |
| 31 | Homa Bay | 3.410303 | 0.386329 |
| 32 | Bomet | 2.631657 | 0.229037 |
| 33 | Nyamira | 1.504872 | 0.071882 |
| 34 | Narok | 6.86177 | 1.454091 |
| 35 | Kisii | 1.67694 | 0.106043 |
| 36 | Murang'a | 2.536662 | 0.206462 |
| 37 | Migori | 3.01159 | 0.257126 |
| 38 | Kiambu | 3.085721 | 0.209126 |
| 39 | Machakos | 5.79204 | 0.505507 |
| 40 | Kajiado | 7.280366 | 1.774001 |
| 41 | Nairobi | 1.615813 | 0.057278 |
| 42 | Makueni | 5.753765 | 0.649311 |
| 43 | Lamu | 11.26831 | 0.501665 |
| 110 | Kilifi | 6.858123 | 1.017021 |
| 111 | Taita Taveta | 5.572639 | 1.395854 |
| 135 | Kwale | 5.64662 | 0.671356 |
| 141 | Mombasa | 1.979026 | 0.018444 |

## References

Berger, J.O., de Oliveira, V., and Sanso, B. (2001). Objective Bayesian Analysis of Spatially Correlated Data. *Journal of the American Statistical Association*, Vol. 96, No. 456, pp. 1361-1374.

Brunsdon, C., Fotheringham, S., and Charlton, M. (1996). Geographically weighted regression: A new method for exploring spatial nonstationarity. *Geographical Analysis*, Vol. 28, *No.* 4

Cox, D.R., Hinkley, D.V. (1974). *Theoretical Statistics*, Chapman & Hall

Cressie, N.A.C. (1993). *Statistics for Spatial Data*, Revised Edition. New York: John Wiley and Sons.

Gamerman, D. (1998). Markov chain Monte Carlo for dynamic generalised linear models. *Biometrika* 85, 215–227.

Gibbons, S. (2002*). Geography, Resources and Primary School Performance*. CEE Discussion Papers 0025, Centre for the Economics of Education, LSE.

Jagero,N. (2013). Poverty Level and Kenya Certificate of Primary Education Performance in Kenya. *British Journal of Humanities and Social Sciences*, Vol 9:1, 18-26.

Maliki, A.B., Ngban, N, and Ibu, J.E. (2009). Analysis of Students' Performance in Junior Secondary School Mathematics Examination in Bayelsa State of Nigeria. *Stud Home Comm Sci*, 3(2): 131-134

Martin, D. and Atkinson, P. (2001). Investigating the Spatial Linkage of Primary School Performance and Catchment characteristics. *Geographical and Environmental Modelling*, Vol. 5, No. 1. 67-83.

Martino, S., Rue, H., 2010. *Implementing Approximate Bayesian Inference using Integrated Nested Laplace Approximation: a manual for the INLA program*. URL: http://www.math.ntnu.no/~hrue/GMRFsim/manual.pdf

Moussavi, M., and McGinn, N. (2009). *A Bayesian Network for School Performance*.

Ogeto, A.J. (2012). *The impact of post-election violence on K.C.P.E pupils' performance in public primary schools in Esise division, Nyamira county, Kenya*. A research project submitted in partial fulfillment of the requirement for the award of a degree of Master of Education, Kenyatta University.

Rue, H. (2007). Approximate Bayesian Inference for Hierarchical Gaussian Markov Random Field Models. *Journal of Statistical Planning and Inference*, Vol. 137, 3177-3192

Rue, H., and Held, L. (2005). *Gaussian Markov Random Fields: Theory and Applications*. Chapman and Hall/CRC, FL.

Rue, H., Martino, S, and Chopin, N. (2009). Approximate Bayesian Inference for Latent Gaussian Models Using Integrated Nested Laplace Approximations. *Journal of the Royal Statistical Society,* Series B 71 (2), 1-35

Saha, G. (2011). Applying logistic regression model to the examination results data. *Journal of Reliability and Statistical Studies,*Vol. 4, Issue 2 . 105-117

Stoffer, D. (1986). Estimation and identification of space-time ARMAX models in the presence of missing data. *Journal of the American Statistical Association*, 81, 762-772

Tobler W., (1970). A computer movie simulating urban growth in the Detroit region. *Economic Geography*, 46(2): 234-240.

Xu, J. (2009). School Location, Student Achievement, and Homework Management Reported by Middle School Students. *The School Community Journal*, Vol. 19, No. 2.

Zimmerman, B. J., & Kitsantas, A. (2005). Homework practices and academic achievement: The mediating role of self-efficacy and perceived responsibility beliefs. *Contemporary Educational Psychology*, *30*, 397-417.