Comparison of Spatial and Conventional Regression Models in Determination of Trachoma Prevalence and Associated Risk Factors

Article · January 2017
DOI: 10.4172/2327-4581.1000170

CITATIONS 0
READS 59

3 authors:

Kirui Kipngetich
Jomo Kenyatta University of Agriculture and …
1 PUBLICATION 0 CITATIONS
SEE PROFILE

Benson Kenduiywo
Jomo Kenyatta University of Agriculture and …
21 PUBLICATIONS 22 CITATIONS
SEE PROFILE

Edward Waithaka
Jomo Kenyatta University of Agriculture and …
17 PUBLICATIONS 31 CITATIONS
SEE PROFILE

Some of the authors of this publication are also working on these related projects:

GIS applications in Kenya View project

Crop type Classification using radar View project
Comparison of Spatial and Conventional Regression Models in Determination of Trachoma Prevalence and Associated Risk Factors

Pius Kipngetich Kirui*, Benson Kipkemboi Kenduuiywo and Edward Hunja Wathithaka

Abstract
Trachoma is a neglected tropical disease and leading infectious cause of blindness, in Kenya it accounts for 19% of blindness. Past research on associated risk factors in Kenya have relied on traditional impact survey data only however non uniform distribution of prevalence in suspected endemic areas despite similar interventions measures calls for the need to include environmental and climatic potential risk factors in modeling trachoma transmission. Our study therefore aims at determining the prevalence of trachoma and its associated risks factors by use of spatial regression models in variable selection, estimation and prediction compared to conventional regression models. Through use of data from trachoma surveys and remotely sensed environmental and climatic data, spatial and non-spatial regression models were implemented. Regression results were then utilized in spatial interpolation using kriging and geographically weighted regression. Rainfall, presence of flies in children’s face, dirty faces of children and aridity were found out to be the significant variables that contributes towards trachoma transmission. Spatial lag model had the least value of akaike information criterion of 385.08 hence performed relatively better compared to the rest of the regressions models. In estimation of prevalence in places where data was not collected, multivariate regression kriging did slightly better than the geographically weighted regression. The study shows that Spatial regression models performs better compared to conventional regression models both in variable selection and in spatial prediction of trachoma prevalence. Among the spatial regressions the significant variables as obtained were similar though spatial lag performed relatively better compared to other regression models in variable selection based on AIC value and R-squared. There was minimal variation between the two spatial interpolation methods.

Keywords
Spatial lag; kriging; Geographical weighted regression

Introduction
Trachoma has been classified by World health organization as tropical neglected dis-ease and a leading infectious cause of blindness. 1.8 million people are visually impaired due to trachoma which has been identified as a health problem in 42 countries [1] Global health Organization, In Kenya it accounts for 19% of the estimated 250,000 blind persons and 7.3 millions of people are at risk of contracting trachoma. Among the regions which were identified as endemic based on the 2004 baseline surveys was Narok [2]. Trachoma transmission cycle involves combination of multiple potential risk factors that through their interaction leads to trachoma transmission. It includes socio-economic, environmental and climatic factors. Given that these factors are spatial in nature spatial distribution of trachoma can be determined based on statistical relationships among the factors and trachoma [3]. Mohebbi et al. [4] in their study notes that by understanding interplay between prevalence of diseases and their associated risk factors provides for the development of statistical models. Mathematical models have previously been used in modeling transmission of infection diseases including trachoma and based on the nature of transmission cycle of trachoma it can easily be modeled Liu et al. [5] hence by use of trachoma incidences data and the suspected potential risk factors that interact within the transmission cycle it is possible to define the relationship among the variables and trachoma by way of mathematical modelling.

Mathematical modeling through regression aims at explaining the relationship among the variables, prediction and estimation based on the established and defined relationship. Results of regression models depends on data type, assumption made structure and form of the model, not well defined mathematical model can lead to unreliable results [6]. In conventional statistics there usually assumption that observations are independent of each other, which might not be the case if we consider the effect of autocorrelation. There exist a level influence among the observations which are closer to each other as pointed out by the work of Tobler et al. [7] that “everything is related to everything else, but near things are more related than distant things”. To avoid getting biased and unreliable results spatial autocorrelation has to be taken into account by use of spatial based statistical models.

Spatial regression models generally vary in the way they account for spatial auto-correlation. Auchincloss et al. [8] in their study classified spatial regression models based on computational complexity, accounting for spatial heterogeneity and explanation of residuals. Spatial error models accounts for spatial autocorrelation in the residuals while spatial lag accounts for autocorrelation in the model with the weights matrix and assumes dependencies exist directly among dependent variable. Spatial durbin model performs a lag effect of the dependent and independent variables [9]. In implementation of spatial regression models weights have to be defined beforehand, the weights is a product neighborhood, defining how and how many neighbor observations interact greatly influences the structure of regression model and has direct impact on the results as noted by Jerrett et al. [10] that is major issue in the assessment of autocorrelation is the selection of a spatial weights which can be explained by the autocorrelation tests. Another way of distinguishing between the spatial models is how the model function is applied over the area under consideration. The spatial autoregressive models can be viewed as global model since they use one function over the whole area Brunsdon et al. [11] proposed geographical weighted regression model (GWR). It is a local model that accounts for heterogeneity
through calculation of coefficients at each measurement location point then interpolates to get a coefficient surface. The need for a spatial regression that accounts for heterogeneity is as pointed out by Smith et al. [12] spatial distribution of trachoma is likely to arise from relationships between risk factors for trachoma at different scales. GWR is best suited to determine the level of influence of each of the potential risk factor and be able to quantify influence of each factor from a coefficient surface. Intervention measures can easily be planned based on the results from GWR as noted out by [5] modeling should be able to guide in intervention measures through accurate estimation and spatial prediction.

The significant risk factors identified can then be used to predict prevalence as argued by Clements et al. [3] that development of spatial based statistical models have led to prediction of disease based on associated risk factors. Through use of spatial interpolation methods such as GWR and multivariate regression kriging trachoma prevalence is estimated. Regression kriging combines a regression of the dependent variable on dependent variables with simple kriging of the regression residuals [13]. Spatial interpolation of prevalence based on established relationship between dependent and independent variables is reinforced by availability of data from the satellite sources.

Our study seeks to identify potential risk factors for trachoma transmission and to estimate prevalence of trachoma based on identified risk factors through use of regression models. Spatial autoregressive models are used for variable selection and to compare their performance with generalized least squares in determination of significant variables. The determined variables are the used in geographical weighted regression to get raster coefficients and to spatially predict prevalence together with regression kriging.

Materials and Methods

Study area

The study is being conducted at Narok County in Kenya. Narok County is one of the counties suspected to be endemic from the impact survey conducted in 2004 Karimurio et al. [14], subsequent surveys have been conducted for the year 2010 and 2014 by ministry of health and operation eyesight universal. Their survey was approved by National Ethical Review Committee [15]. Figure 1 shows the study area.

Data Collection

The study uses data from the 2010 and 2014 survey reports. The survey report data sets resolution is at sub location level and it includes data on active trachoma, facial cleanliness, hygiene and environmental sanitation. The environmental and climatic data was acquired from satellite sources and it includes rainfall, temperature, altitude and aridity. Rainfall was retrieved online courtesy of Climate Hazards Group (CHG) [16]. Aridity was from the work of Deblauwe et al. [17] it consist of yearly sum of precipitation divided by the annual Potential Evapotranspiration provided by MODIS sensor. The SRTM data product was retrieved from the online Data Pool, courtesy of the NASA Land Processes Distributed Active Archive Center (LPDAAC), USGS/Earth Resources Observation and Science (EROS) Center, Sioux Falls, South Dakota [18].

Methods

Figure 2 shows general work flow that was adopted in carrying this study. It involved combining datasets from surveys and satellite sources into one database. Cluster points were used to extract datasets from the satellite sources to obtain a point shapefile consisting of all the factors as attributes of the shape file. R software was used to implement the regression models [19]. Table 1 shows the structure of regression models that were used in this study. Where y is a N × 1 vector of observations of the dependent variable, X is a N × K matrix of observations of the explanatory variables, β is a K × 1 vector of regression coefficients, ε is N × 1 vector of spatially auto correlated error terms, W represent weighted matrix, λ is the autoregressive

![Figure 1: Study area.](image1)

**Table 1: Regression models.**

<table>
<thead>
<tr>
<th>Regression type</th>
<th>Form</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ordinary least squares model</td>
<td>$y = X\beta + \epsilon$</td>
</tr>
<tr>
<td>Spatial lag model</td>
<td>$y = \pi W y + X\beta + \epsilon$</td>
</tr>
<tr>
<td>Spatial Durbin model</td>
<td>$y = \pi W y + X\beta + \epsilon + U$</td>
</tr>
<tr>
<td>Spatial error model</td>
<td>$y = X\beta + \epsilon, \epsilon = \lambda W \epsilon + u$</td>
</tr>
</tbody>
</table>
coefficient \( u \), \( U \) is an error term, and \( \rho \) is a scalar parameter that indicates the effect of the dependent variable in the neighbors.

Generalized least square regression was implemented using nlme package Pinheiro et al. [20] and spdep Gls package was used to implement the spatial regression models [21]. The structure of spatial regression models as shown in Table 1 contains weights which is product of neighborhood function. K neighborlist were created then weights were assigned to defined neighborlist. 3 to 10 k-neighbors were chosen to create several weight matrices. Several Spatial lag models were used to obtain optimal weight matrix. Weight matrix of the model having least AIC value being adopted for implementation of spatial regressions models as pointed out by Gerkman et al. [22] weight matrix defines the assumed spatial structure and interaction of variables in the model.

Four regression models were used to determine and compare significant variables from each of the model based on the p-value. Backward stepwise selection method was used in determining significant variables based on p-value. P-value is a conditional probability against null hypothesis proposed by Fisher et al. [23], null hypothesis for this study is that no relationship between variables and trachoma transmission. All the variables were used in modeling, the variables that had p value of less or equal to 0.05 were retained and considered significant and deemed to have a role in transmission of trachoma.

The significant variables as obtained from the optimum regression model based on Akaike information criterion (AIC) were then utilized in spatial prediction using regression kriging and geographical weighted regression. Regression kriging was implemented in R software using gstat package by Graler et al. [24]. Geographical weighted regression was implemented using ArcGIS [25]. In specifying the type of Gaussian interpolation, adaptive kernel was used as pointed out by Hu et al. [26], adaptive kernel scheme is good in solving local regression model to ensure maximum number of features is used in interpolation.

**Results**

**Neighbor list selection**

Starting with few k neighborlist results to high AIC values, as the number of k-neighbors increases AIC values decreases. AIC values stabilizes between 7 to 9 neighbors where it starts to increase sharply Figure 3 shows a graph of AIC values against the neighbor list.

**Table 2: Significant variables.**

<table>
<thead>
<tr>
<th>Model name</th>
<th>Spatial lag</th>
<th>Spatial Durbin</th>
<th>Spatial error</th>
<th>Gls</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dirty face</td>
<td>2.75e-05</td>
<td>1.86e-06</td>
<td>6.26e-05</td>
<td>2.87e-02</td>
</tr>
<tr>
<td>Flies</td>
<td>3.35e-04</td>
<td>6.02e-03</td>
<td>1.83e-4</td>
<td>7.58e-02</td>
</tr>
<tr>
<td>Rain</td>
<td>9.92e-06</td>
<td>1.38e-05</td>
<td>25.9e5</td>
<td>1.58e-08</td>
</tr>
<tr>
<td>Altitude</td>
<td>1.22e-03</td>
<td>1.40e-03</td>
<td>3.57e-02</td>
<td>2.18e-05</td>
</tr>
<tr>
<td>Aridity</td>
<td>5.27-02</td>
<td>4.8e-03</td>
<td>1.33e-02</td>
<td>4.06e-02</td>
</tr>
<tr>
<td>Lambda</td>
<td>4.85e-06</td>
<td>6.81e-06</td>
<td>2.45e02</td>
<td>2.45e02</td>
</tr>
</tbody>
</table>

**Significant variables**

Summary characteristics of the significant variables based on the p-values is as shown in Table 2. Rainfall, presence of flies in children’s face, dirty faces of children and aridity were found out to be the significant variables that contributes towards trachoma transmission.

**Regression coefficients**

The effect on unit change of each of the significant variables is as shown in Figure 4. For flies it ranges from 0.01–0.41 increasing towards south East. Dirty face varies from 0.01–0.41 and follows the same trend with flies only that it increases uniformly towards the south. The effect of aridity follows different pattern as it shows that some areas are negatively correlated while other areas are positively. Rainfall effect on transmission varies from 0.01 to 0.02 and it increases southwards.

**Model performance**

To evaluate the performance of regression models coefficient of multiple determination for multiple regression (R2), scatter plots of predicted against observed values and AIC were used. Generalized least squares had highest AIC value of 400.12 and the least R2 value of 0.772. Spatial lag model had the least AIC value of 385.08 with little variation of R2 among the spatial regression models as shown in Figure 5. High discrepancy between observed and predicted values was noted in GLS model while spatial regression models exhibited low discrepancy. As shown in Figure 5 spatial regression models performed relatively better compared to generalized least squares model.

**Mapping Residuals**

Spatial regression models results were checked if any spatial autocorrelation still exists after being accounted for this was achieved through mapping of residuals Figure 6.
Figure 7 shows prevalence as predicted using regression kriging and GWR. Trachoma prevalence was found to increase southwards. There was minimal discrepancy both spatially and quantitatively as predicted by both methods. R-squared for GWR was 0.76 while for regression kriging was found to be 0.79.

Discussion

Weight matrix defines the structure of spatial dependence however challenge is in defining and understanding the structure beforehand. As shown in Figure 3 weight matrix with 7 k neighborlist yielded the least AIC value of 385.08, additionally model performance stabilized between 6 to 8 neighbors. It is at this point that spatial dependence is well explained which translates to better model performance. Increasing or decreasing the number of neighbors in either direction affects the structure of spatial dependence this explains the sharp decline in model performance on either side.

The significant variables as obtained from regression models shown in Table 2 supports existing expert views on the role played by reported variables in trachoma transmission as reported by Vinke et al. [12, 27]. Spatially varying coefficient estimates of the significant variables as shown in Figure 4 can play role in guiding intervention measures, it show factors and regions to concentrate on. For instance aridity coefficient estimate exhibits both negative and positive correlation which can be attributed to decreasing effect of aridity in trachoma transmission in areas where the coefficient is negative. The effect of dirty face is more pronounced in southern part which may be attributed to short- age of water that restricts the use of water, which is also supported by flies coefficient estimate that follows the same trend with dirty face.

Model comparisons of the three spatial models was to determine optimum model and for purposes of validation. Scatter plots are for visual interpretations between predicted and observed values, the trend line represents ideal situation where predicted values matches the observed values implying no residuals. spatial regression models as shown in Figure 5 had minimum deviation between predicted and observed values. AIC uses likelihood function in conjunction with the number of independent variables to compare the models, low AIC values indicates better relative performance while high R squared values closer to one indicates how well the model has accounted for the variance. The minimal variation between regression kriging and GWR in estimation of prevalence is due to the fact that both accounts for spatial autocorrelation during interpolation. Few cases of clustering by GWR can be attributed to GWR being a local model captures local variability in prevalence estimation.

Conclusion

The choice of weights matrices greatly influences the results of spatial regression models and care must be taken in deciding structure of weight matrix that best describes the level of influence among the observations under consideration otherwise the results will be misleading.

Spatial regression models performed better compared to conventional regression models both in variable selection and in spatial prediction of trachoma prevalence. Among the spatial regressions the significant variables as obtained were similar though spatial lag performed relatively better compared to other regression models in variable selection based on AIC value and R -squared. There was minimal variation between the two spatial interpolation methods.
By utilizing spatial regression models such geographical weighted regression it can be noted that the role of spatial regression modeling in epidemiology should not be limited to explaining the relationship between the variables and in prediction but should extend as and play a role in guiding the intervention measures through spatial variability of the risk factors, interrelationships between the variables and the magnitude.

Acknowledgement

We wish to extend our appreciation to Kenya national trachoma control office for provision of survey reports and coordinates of cluster points.

References

18. NASA, Data Pool (2014) Land Processes distributed active archive center, USA.
25. ESRI, ArcGIS Release 10.3 Environmental Systems Research Institute, Redlands, USA.

Author Affiliation

Department of Geomatic Engineering and Geospatial Information System, Jomo Kenyatta University of Agriculture and Technology, Nairobi, Kenya

Submit your next manuscript and get advantages of SciTechnol submissions

- 80 Journals
- 21 Day rapid review process
- 3000 Editorial team
- 5 Million readers
- More than 5000
- Quality and quick review processing through Editorial Manager System

Submit your next manuscript at ● www.scitechnol.com/submission