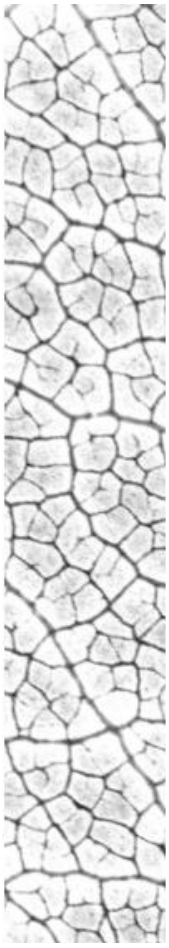

EurAsian Journal of BioSciences
Eurasia J Biosci 14, 4415-4420 (2020)

Application of geographically weighted regression to assess risk factors for water pollution related human diseases

R. Sasikumar [1*], S. Raguraman [1]

[1] Department of Statistics, Manonmaniam Sundaranar University, INDIA
*Corresponding author: sasikumarmsu@gmail.com

Abstract
Water is essential for survival. Human health may be affected directly or indirectly by the ingestion of contaminated water and by the use of polluted water for purposes of personal hygiene. The water related diseases data and water pollution data were analysed with ordinary linear regression and geographically weighted regression by using R software. The results of the analysis show that geographically weighted regression model can be used to geographically differentiate the relationships of water related diseases with water pollutants. This paper studied the factors affecting human health due to drinking water quality in Tirunelveli district, Tamil Nadu.

Keywords: water pollution, waterborne diseases, water related diseases, ordinary linear regression, geographically weighted regression

Sasikumar R, Raguraman S (2020) Application of geographically weighted regression to assess risk factors for water pollution related human diseases. Eurasia J Biosci 14: 4415-4420.

INTRODUCTION

All forms of life depend on water. About 70% of human body weight is water. The human brain is 85% water, blood is 82% water and lungs are 90% water. Today most of the people in India do not have access to safe drinking water. Most of the water resources are polluted with untreated/partially treated wastes from industry, domestic sewage and fertilizer/pesticide run off from agriculture fields. Much of the ill-health in the underdeveloped countries is largely due to lack of safe drinking water. There can be no state of positive community health and well-being without a safe water supply. Health problems can arise by ingesting contaminated water (drinking the water, eating food that has lived in the water) as well as through airborne exposures from materials that can outgas during showering, bathing, swimming or cooking. It can also be absorbed through the skin Brunsdon et. al. (1999).

These diseases can be contracted directly or indirectly through exposure to infected vectors, but water is the primary medium for their spreading. Intestinal (enteric) diseases are generally infectious, transmitted through fecal waste, pathogens are bacteria, viruses, protozoa or parasitic worms capable of producing diseases. Pathogens are often found in faces of infected persons. Infectious diseases are those where the pathogen is capable of entering, surviving and multiplying in the host. Such diseases are more common in areas with poor sanitary conditions. Pathogens can travel through water and directly affect persons exposed to the water. Diarrhoea, hepatitis, cholera and typhoid are the more common waterborne diseases and they occur more frequently in tropical regions of the world. In addition, several chemicals, both human-made and existing dissolve in water, polluting the water and producing disease among those exposed Brunsdon et. al. (1999).

Human Activities in Tambiraparani River

The Tambiraparani river is being polluted by human bathing, cloth washing, textile-mill effluent, sewage, utensil cleaning, vehicle cleaning and cattle washing activities of human beings. In Tirunelveli Corporation where the Tambiraparani river flows through the city, sewage enters the river without any treatment. Human inhabitation is seen on both sides of the river.

These unhealthy human activities reduce the amount of pure fresh water that is available for such necessities as drinking and cleaning and for recreation activities. Being a universal solvent water receives contamination by the material with which it comes into physical contact. This water becomes unfit for human consumption.

Common Diseases due to Water Pollution

Waterborne diseases are infectious diseases that are spread primarily through contaminated water. Major

Received: July 2019
Accepted: April 2020
Printed: October 2020

water borne diseases are classified as follows. Typhoid (enteric fever), Cholera, Leprosy, Gastro Enteritis, Cerebrospinal Meningitis, Diptheria, Whooping Cough, Tuberculosis, Fever, Viral Encephalitis, Virus Fever, Common Cold, Hepatitis A (Jaundice), Mumps, Measles, Chickenpox, Diarrhoea, Dysentery, Stone-formation, Scabies, Malaria, Tetanus, Asthma, Depression, Madness and Snake bites.

Though Tirunelveli Corporation applies many methods to purify drinking water, diseases like, diarrhoea, typhoid, infective hepatitis A, gastro enteritis, cholera and bacillary dysentery are prevalent in Tirunelveli Corporation. These type of diseases occur only when drinking water is polluted. Malaria, Scabies and Viral (Japanese) encephalitis, also affect the Tirunelveli Corporation public which is water related environmental diseases.

This paper is to describe the role of water in human health, identify common waterborne and water related diseases and discuss some of the monitoring efforts due to drinking water quality in Tirunelveli district of Tamil Nadu.

Review of Related Work

Frank Curriero et. al., (2001) studied the relationship between precipitation and waterborne diseases, using the complete database of all reported waterborne disease outbreaks in the United States from 1948 to 1994. Fahad et. al., (2006) proposed the morbidity of typhoid fever was highest in Asia with 93% of global episodes occurring in this region. Facility-based surveillance missed the significant proportion of potential cases raising the need for even more extensive surveillance systems. Dipika Sur et.al.,(2007) compared and contrast risk factors of typhoid and paratyphoid fever and on an individual level as well for Kolkata areas with increased incidence. The comparison of risk factors of populations living in high versus low risk areas is statistically very powerful this methodology holds promise to detect risk factors associated with diseases using geographic information systems.

Claudia et. al., (2008) studied the model with an application to time series data for two water-borne diseases, leptospirosis and cholera, and their seasonal response to rainfall. The quantitative description of this kind of threshold behaviour was more general application to predict the response of ecosystems and human health to climate change. Gordon et. al., (2009) examined the relationship between rainfall and outbreaks of drinking water related disease. For drinking water provides the results emphasize that specific climatic conditions have increased the risk of outbreaks occurring and suggest that interventions focused on periods of low rainfall as well as following heavy rain might be useful in formulating Water Safety Plans. Dewan et. al., (2013) studied typhoid infection on the Dhaka Metropolitan Area of Bangladesh. Spatial and environmental factors were used to identify possible causal factor for typhoid incidences has studied. In addition to these factors, other variables such as population density can be used to examine the factors that are most responsible at the local level.

Data Sources

Hospital Data

Hospital admission data in Tirunelveli district from January 2012 to December 2016 for waterborne and water related diseases were obtained from Tirunelveli Government Hospital. Nine diagnostic categories were Whooping Cough, Measles, Tuberculosis, Typhoid, Viral Fever, Hepatitis – A, Viral Encephalitis, Dysentery and Gastro Enteritis.

Study Area

Tirunelveli also known as Nellai and historically as Tinnevelly is a city in the south India state of Tamil Nadu. The city is located on the west bank of the Thamirabarani river its twin city Palayamkottai is on the east bank. Tirunelveli is located at 8.73^{o} N, 77.7^{o} E and its average elevation is 47 metres (154 ft). The Tamirabarani river divides the city into the Tirunelveli quarter and the Palayamkottai area. The river (with its tributaries such as the Chittar) is the major source of irrigation and is fed by the northeast and southwest monsoons.

Environmental Data

Tamirabarani river pollution data were obtained by **www.cpcb-gems/minars water quality criteria.in**. These observations of water pollution data such as Dissolved Oxygen (DO), Biochemical Oxygen Demand (BOD), Chloride (Cl), Sulphate (SO_4), Nitrate (NO_3), Faecal Coliform (FC), Total Coliform (TC), Total Hardness as $CACO_3$ (TH), potential of Hydrogen (PH). Water pollution data measured by (mg/L).

METHODOLOGY

The linear regression estimates of a parameter β that links the explanatory variables to the response variable. However, when this technique is applied to spatial data, some issues concerning the stationarity of these parameters over the space come out. In normal regression, it is generally assumed that the modelling relationship holds everywhere in the study area that is, the regression parameters are whole map statistics. In many situations this is not the case, however, as mapping the residuals the difference between the observed and predicted data may reveal. The realization in the statistical and geographical sciences that a relationship between an explanatory variable and a response variable in a linear regression model is not always constant across a study area has led to the development of regression models allowing for spatially varying coefficients. Many different solutions have been proposed for dealing with spatial variation in the

relationship. One of them, developed by Brunsdon et al., (1996) has been labelled Geographically Weighted Regression (GWR), which provides an elegant and easily grasped means of modelling such relationships by subtly incorporating the spatial characteristics of data via allowing regression coefficients to depend on some covariates such as longitude and latitude of the meteorological stations. Specifically, it is a nonparametric model of spatial drift that relies on a sequence of locally linear regressions to produce estimates for every point in space by using a sub sample of data information from nearby observations. That is to say, this technique allows the modelling of relationships that vary over space by introducing distance-based weights to provide estimates β_{ki} for each variable k and each geographical location i. Thus the spatial variation of regression relationship can be effectively analysed and the inherent disciplines of spatial data by the estimated coefficients over different locations can be better understood.

An Ordinary Linear Regression (OLR) model can be expressed by,

$$y_i = \beta_0 + \sum_{j=1}^{m} \beta_j x_{ij} + \varepsilon_i, i = 1,2,\dots,n \quad (1)$$

where y_i, i=1,2,...,n are the observation of the response variable y, β_j(m=1,2,...,p) represents the regression coefficients, x_{ij} is the i^{th} value of the explanatory variable x_j, and ε_i are normally distributed error terms with zero mean and constant variance.

In GWR model, the global regression coefficients are replaced by local parameters,

$$y_i = \beta_0(u_i, v_i) + \sum_{j=1}^{m} \beta_j(u_i, v_i) x_{ij} + \varepsilon_i, i = 1,2,\dots,n \quad (2)$$

where (u_i,vi) denotes the longitude and latitude coordinates of the i^{th} meteorological station, $(y_i; x_{i1}, x_{i2}, \dots, x_{ip})$ represent the observed value of the response Y and explanatory variables $X_1,X_2,...,X_p$ at (u_i,v_i), $\beta_0(u_i,v_i)$ is the intercept and β_j (u_i, v_i) (j=1,2,...,p) are p unknown coefficient functions of spatial locations, which represent the strength and type of relationship that the j^{th} explanatory variable X_j has to the response variable Y. Additionally, $\varepsilon_1,\varepsilon_2,...,\varepsilon_n$ are error terms which are generally assumed to be independent and identically distributed variables with mean 0 and common variance σ^2. It is worth noticing that the OLR model is actually a special case of the GWR model where $\beta_j(u_i, v_i)$ are constant for all i=1,2,...,n.

The coefficient function vector $\hat{\beta}(u_i, v_i)$ for the i^{th} observation in GWR can be estimated via the locally weighted least square procedure Brunsdon et. al. (1999) as,

$$\hat{\beta}(u_i, v_i) = \left(\hat{\beta}_0(u_i, v_i), \hat{\beta}_1(u_i, v_i), \dots, \hat{\beta}_p(u_i, v_i)\right)^T$$
$$\hat{\beta}(u_i, v_i) = (X^T W(i) X)^{-1} X^T W(i) Y, i = 1,2,\dots,n \quad (3)$$
$$\hat{\beta}_j(u, v) = \left(\hat{\beta}_j(u_1, v_1), \hat{\beta}_j(u_2, v_2), \dots, \hat{\beta}_j(u_n, v_n)\right)^T, j = 0,1,2,\dots,p$$

where

$$X = \begin{pmatrix} x_1^T \\ x_2^T \\ \vdots \\ x_n^T \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & \cdots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{pmatrix}, \; Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \quad (4)$$

$$W(i) = diag[K_h(d_{i1}), K_h(d_{i2}), \dots, K_h(d_{in})] \quad (5)$$

is a diagonal weight matrix, ensuring that observations near to the location have greater influence than those far away. Here, d_{ij} denotes the distance between two observed locations (u_i,v_i) and (u_j,v_j), which can be calculated as,

$$d_{ij} = R\arccos\left(\sin v_i \sin v_j + \cos v_i \cos v_j \cos(u_i - u_j)\right) \quad (6)$$

where R is the earth radius. In equation (5), K_h (.) = 1/hK (. / h) with K(.) being Gaussian Kernal function,

$$K(t) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2} t^2\right) \quad (7)$$

and h is the bandwidth which can be estimated by some data-driven procedures such as the Cross Validation (CV) method Desmet and Gijbels (2009), the Generalized Cross Validation (GCV) procedure Fotheringham et. al. (2002), or the corrected Akaike Information Criterion (AIC) Hurvich et. al. (1998). In this paper, the adaptive kernel with AIC estimated bandwidth was chosen. The adaptive kernel was chosen because the distribution was inhomogeneous in the study area.

Although GWR is very appealing in analysing spatial non-stationarity, from the statistical viewpoint, two critical questions still remain. One is the goodness of fit test, that is, OLR model is compared to a GWR model to see which one provides the best fit. Usually, a GWR model can fit a given data set better than an OLR model. However, the simpler a model, the easier it can be applied and interpreted in practice. If a GWR model does not perform significantly better than an OLR model, it means that there is no significant drift in any of the model parameters. Thus, we will prefer an OLR model in practice. On the other hand, if a GWR model significantly outperforms an OLR model, we will be concerned with the second question, that is, whether each coefficient function estimate $\hat{\beta}_j(u, v)(j = 1,2,\dots,p)$ exhibits significant spatial variation over the studied area Leung et. al (2000), Brunsdon et. al. (1999). If the answer to this question is positive, the characteristics of the data will be investigated in more details.

The data were analysed with OLR and GWR by using GWmodel and spgwr packages in R software.

RESULTS AND DISCUSSION

There were 180 hospital admissions with water borne and water related diseases during the month of January 2010 to December 2016. A summary of the variables in both OLR and GWR models are shown in **Table 1**.

The adjusted coefficient of determination (Adjusted R^2) was used for comparing OLR and GWR models. Akaike Information Criterion (AIC) generated for OLR

Table 1. Summary of Dependent and Independent Variables used in OLR and GWR

Dependent Variable	Independent Variable
Water borne and Water related Diseases (Whooping Cough, Measles, Tuberculosis, Typhoid, Viral Fever, Hepatitis – A, Viral Encephalitis, Dysentery and Gastro Enteritis)	**Water Pollutants** (Dissolved Oxygen (DO), Biochemical Oxygen Demand (BOD), Chloride (Cl), Sulphate (SO_4), Nitrate (NO_3), Faecal Coliform (FC), Total Coliform (TC), Total Hardness as $CACO_3$ (TH), Potential of Hydrogen (PH))

Table 2. Ordinary Linear Regression (OLR) Results

Parameter	Estimated Value	Std. Error	Adjusted R^2	AIC	P – Value
Whooping Cough	202.876	5.808	0.04716	1108.675	0.038
Measles	198.111	6.926	0.03137	1220.089	0.096
Tuberculosis	367.483	7.556	0.48620	1249.899	0.000
Typhoid	68.998	7.556	0.48620	1249.899	0.000
Viral Fever	156.979	5.293	0.00322	1952.037	0.495
Hepatitis A	146.930	5.293	0.00322	1952.124	0.495
Viral Encephalitis	104.600	7.097	0.04572	1228.868	0.041
Dysentery	106.467	7.393	0.00468	2072.270	0.362
Gastro Enteritis	123.100	2.208	0.07611	1637.213	0.005

Table 3. Geographically Weighted Regression (GWR) Results

Parameter	Bandwidth	Adjusted R^2	AIC
Whooping Cough	0.1948356	0.090315	1096.065
Measles	0.1948356	0.074805	1207.479
Tuberculosis	0.1162082	0.516212	1236.733
Typhoid	0.1162082	0.551322	1234.390
Viral Fever	0.1948356	0.042899	1939.427
Hepatitis A	0.1948355	0.059962	1192.215
Viral Encephalitis	0.1948356	0.088761	1216.258
Dysentery	0.1948356	0.050630	2059.660
Gastro Enteritis	0.1948113	0.119131	1624.603

and corrected Akaike Information Criterion (AIC) calculated for GWR were also used for comparison.

OLR Model

The results of applying OLR showed that holding the variable of tuberculosis estimated value increase is significantly associated with 367.483 in **Table 2**. However, this global regression model AIC value is 1249.899 and adjusted R^2 value is 0.48620. The variable of typhoid estimated value decreasing is significantly associated with 68.998, this AIC value is 1249.899 and adjusted R^2 value is 0.48620.

We further examined the residuals of the OLR model and found the residuals had positive spatial autocorrelation ($P < 0.01$). Since the existence of dependent residuals the assumptions of OLR estimation, we employed a GWR model to fit the data.

GWR Model and Spatial Variation

The summary of GWR are listed in **Table 3** and showed the GWR was more suitable than the OLR model since GWR could 100 percent of the model variation with decreased AIC. All parameter variables of AIC values are decreasing and adjusted R^2 values are increasing in GWR model.

The scatter plot of the GWR coefficients suggested **multicollinearity was not serious shown in Fig. 1.**

CONCLUSION

This study provides further indications that the relationships of water borne, water related diseases and water pollutants were spatially non-stationary in Tirunelveli District. In regression maps, it is clear that the water borne, water related diseases and water pollutants were different in the study area. We used, GWR since the conventional regression, OLR cannot discriminate the spatial variation in relationships if geographical nonstationary exists. The results of adjusted R^2 and AIC all indicated GWR was a better model to explain this dataset.

ACKNOWLEDGEMENT

The authors acknowledge University Grants Commission for providing the infrastructure facility under the scheme of SAP (DRS-II). The second author acknowledge UGC for financial support to carry out this research work under Basic Science Research Fellowship.

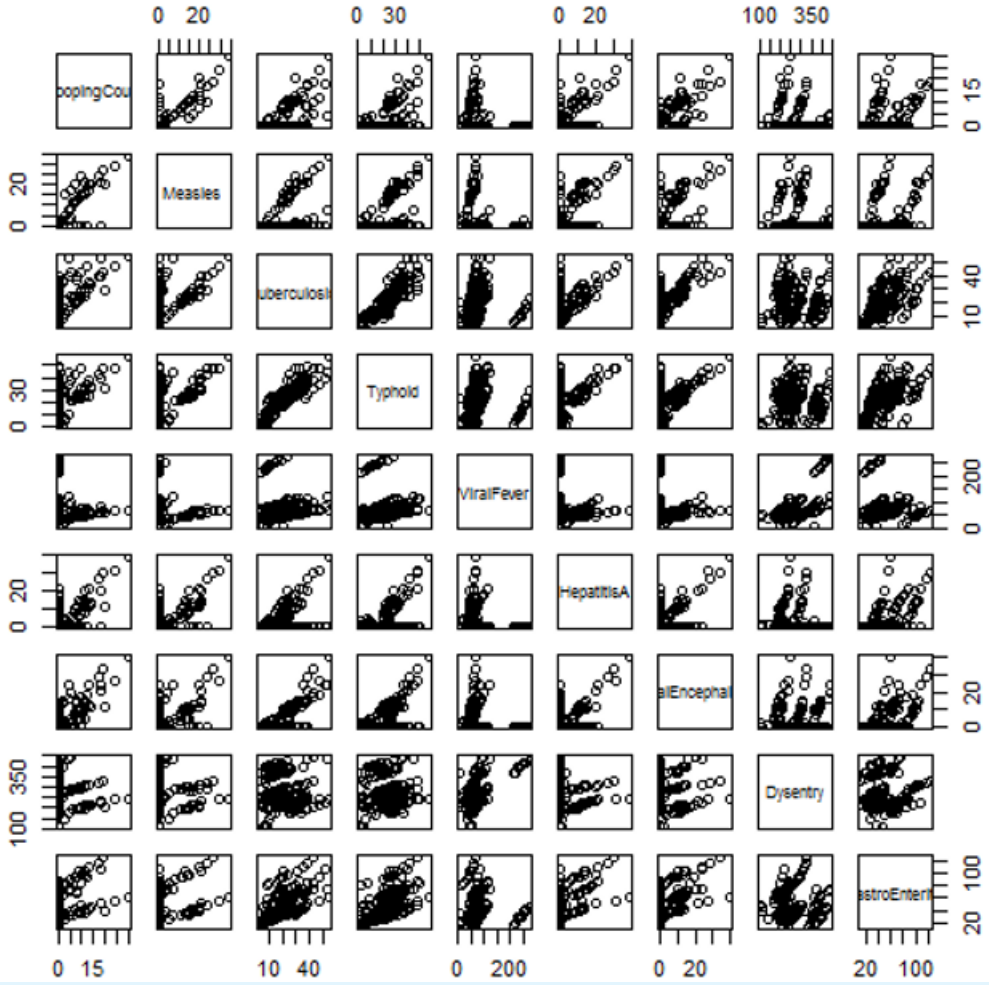

Fig. 1. Scatter Plot of GWR Coefficients of Water Pollutants and Water related diseases. The dashed lines were the levels of the OLR estimators

REFERENCES

Brunsdon, C, Aitkin, M, Fotheringham, S and Charlton, M. (1999). A comparison of random coefficient modelling and geographically weighted regression for spatially non-stationary regression problems, Geographical and Environmental Modelling, 3(1), 47–62.

Brunsdon, C, Fotheringham, A. S. and Charlton, M. (1999). Some notes on Parametric Significance tests for Geographically Weighted Regression, Journal of Regional Science, 39(3), 497-524.

Brunsdon, C, Fotheringham, A. S. and Charlton, M. E. (1996). Geographically weighted regression: a method for exploring spatial nonstationarity, Geographical Analysis, 28(4), 281-298.

Claudia, T.C, Subhash, L, Mercedes, P, Bouma, and Albert, I. K. (2008). A Stochastic Model for Ecological Systems with Strong nonlinear response to Environmental Drivers: Application to two Water-borne Diseases, Journal of The Royal Society Interface, 5, 247-252.

Desmet, L, and Gijbels, I. (2009). Local linear fitting and improved estimation near peaks, The Canadian Journal of Statistics, 37(3), 453–475.

Dewan, A. M. Corner, R, Hashizume, M and Ongee, E. T. (2013). Typhoid Fever and its Association with Environmental Factors in the Dhaka Metropolitan Area of Bangladesh: A Spatial and Time Series Approach, PLOS Neglected Tropical Diseases, 7(1), 1-14.

Dipika, S, Mohammad, A, Lorenz, V. S, Byonkesh, M, Jacqueline, L. D, Camilo, J. A. John DC and Sujit KB (2007). Comparisons of Predictors for Typhoid and Paratyphoid fever in Kolkata, India, BMC Public Health, 7, 289-298.

Fahad JS, Fauziah R, Rumina H, Syed QN and Zulfiqar AB (2006). Typhoid Fever in Children - Some Epidemiological Considerations from Karachi, Pakistan, International Journal of Infectious Diseases, Vol. 10, pp.215-222.

Fotheringham, A, Brunsdon, C and Charlton, M. (2002). Geographically Weighted Regression: The Analysis of Spatially Varying Relationships, Wiley, Chichester, UK.

Frank, C. C, Jonathan, A. P, Joan, B. R and Subhash, L. (2001). The Association between Extreme Precipitation and Water Borne Disease Outbreaks in the United States, 1948-1994, American Journal of Public Health, 91(8), 1194-1199.

Gordon, N, Chris, L, Nima, A, Neville, Q. V. and Andre, C. (2009). Rainfall and Outbreaks of Drinking Water related Disease and in England and Wales, Journal of Water and Health, 7(1), 1-8.

Hurvich, C. M, Simonoff, J. S. and Tsai, C. L. (1998). Smoothing parameter selection in nonparametric regression using an improved Akaike information criterion, Journal of the Royal Statistical Society B, 60(2), 271–293.

Leung, Y, Mei, C. L. and Zhang, W. X. (2000). Statistical tests for spatial non stationarity based on the geographically weighted regression model, Environment and Planning A, 32, 9-32.

Ray, M. M. (2010). Environmental Epidemiology Principles and Methods, Jones and Bartlett India Private Limited, New Delhi.