Thailand Statistician
October 2020; 18(4): 429-438
http://statassoc.or.th
Contributed paper

Analysis of Water Quality by Using Spatial Graph Theory and Metamodelling

Charan Kumar Ganteda*[a] and Gurram Shobhalatha [b]
[a] Department of Mathematics, Koneru Lakshmaiah Education Foundation, K L Deemed to be University, Vaddeswaram, India.
[b] Department of Mathematics, Srikrishnadevaraya University, Anantapur, India.
*Corresponding author; e-mail: charankumarganteda@gmail.com

Received: 29 January 2019
Revised: 21 April 2019
Accepted: 28 August 2019

Abstract

In this paper, we made an attempt to integrate the use of advanced technology, remote sensing and geographic information system (GIS) software by linking with the water quality data to create the spatial distribution maps for identification of water quality stretch zones and its impact by considering spatial statistics, spatial regression, simulation and spatial graph theory. We mainly focused how statistics, simulation and graph theory can be used to provide insights into better understand the various parameters to the effect of water quality. In the usual regression, it has shown less variation but in the simulated graph the high impact of water quality by different parameters exhibited. It will be helpful to take the precautionary steps for better use of resources. From the analysis, we observed that possible interactions are with variable NA, K, F and overall lack of fit test is significant. R^2 is quite impressive and residuals appear to exhibit the cyclical pattern about the regression line.

Keywords: Regression analysis, spatial regression, remote sensing and GIS, simulation.

1. Introduction

In recent years, there has been a growing concern, in India, about the increased deterioration of the country's environment. Water quality testing is an important part of environmental monitoring. According to the WHO, statistical analysis shows that majority of human diseases born from the improper management of water quality. The integration of graph theory, statistical tools and remote sensing with geographic information system (GIS) provides a scientific platform for developing an integrated database among all the different entities involved in environmental planning and management activities. Testing of water quality is an important part of environmental monitoring. When water quality is poor, it affects the entire ecosystem. It is affected by various parameters which can be studied with advancements in remote sensing and GIS. A good amount of experimental work done by Asadi et al.(2005) useful for identifying optimum process parameters to identify the quality of water. Their test results can be utilized to validate the graphical and metamodelling techniques which will be useful to determine the most effected parameters of water quality. Here we consider 11

parameters and identifying the factors which are highly affecting the water quality index and more emphasis on the groundwater quality, sources of ground water contamination, variation of groundwater quality and its spatial distribution, various effects of poor quality of groundwater, impacts of land use/land cover changes on quality of ground water. Many human activities such as urban development, industrial processing, agriculture, chemical spills and even individual household septic systems cause significant ground-water contamination in areas that previously had clean, potable ground water. Ground-water contamination can disperse over a wide area or migrate very deep underground. Often, many tons of overlying soil, sediment or rock hide the exact location of the contamination and present a substantial physical barrier to clean up efforts. As the ground water moves, it often contaminates the earth materials, it passes through which, increases the volume of material that needs to be cleaned. The cost and technical difficulty of removing the contamination often multiplies over time as the contamination spreads out or migrates deeper.

Kobayashi (1997) provided generalization on a spatial graph theory. Gyananath et al. (2001) assessed environmental parameter on ground water quality. Structural landscape connectivity developed by using spatial graph algorithms (Fall et al. 2007, Dale and Fortin 2010). Spatial regression can be classified into three main categories depending on how spatial effects are modeled (Dormann et al. 2007, Beale et al. 2010): i) space included in covariate predictors ii) space included in error term iii) spatial effects in the response or explanatory variables are replaced by transforming the original data. Barton (2015) developed simulation metamodelling. Erica et al. (2016) developed spatial graphs to intrinsic knotting and linking results. Kamaldeep et al. (2017) studied spatial modeling of urban road traffic using graph theory. Spatial analysis is very much useful in analyzing the data related to geography. Animal movements can be studied by the studies ethological; land scape by ecology-population dynamics and biogeography. Epidemiology contributed with early work on spatial graphing an outbreak disease and with location studies for health care delivery. Spatial graph provides a graph in the 3-dimensional Euclidean space R^3 or the 3-sphere. For a graph G we take an embedding $f : G \to R^3$ then the image $\overline{G} = f(G)$ is called a spatial graph of G. It is a generalization of knot and link. Several parameters are influenced the quality of water index.

In this regard, we identify the process parameters which influence the water quality and identifying the maximum influenced factors using statistical and graphical approaches. In this connection, we use the advanced technology of remote sensing and GIS software linking with the water quality data to create the spatial distribution graphs for identification of water quality stretch zones and its impact and we apply the concept of metamodelling to get the residuals from the data and observed that the large residuals can be removed by applying the simulation and getting the accurate result. The spatial graph theory, regression and simulation will help us to give the accurate result when more variables are affecting the water quality index and it is possible to identify the areas which have low quality of water index and unfit for human consumption. From the analysis we get, out of 300 areas 26 observations showed the large standardized residual and 8 observations have larger influence in effecting the water quality index and also observed that possible interactions are with variable NA, K, F and overall lack of fit test is significant. R^2 is quite impressive and residuals appear to exhibit the cyclical pattern about the regression line. In this paper, we mainly focused on integration of the spatial and attribute data base to develop metamodelling for knowing the water quality impact by using spatial graph theory and spatial regression.

2. Methodology

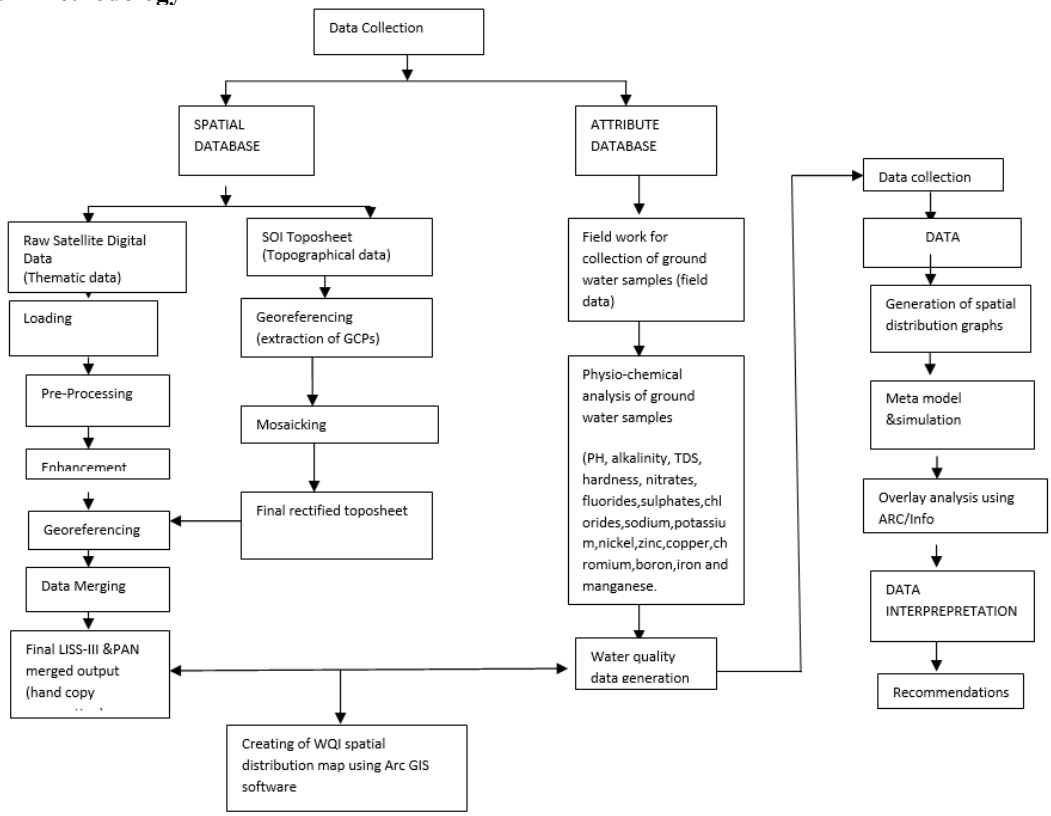

Figure 1 Step by step procedure of metamodelling

Figure 1 gives the complete description about the methodology which we follow to give the conclusions in a systematic manner. First, we collect the data in our selected study region of Hyderabad ground water. We create spatial data base with the help of ARC/INFOGIS/Geoda software developed by Anselin (2006), after that we examine the ground water quality by considering the physic-chemical characteristics of our study region. Next, we identify the parameters and stretch zones which are more influencing the index and showing variations by spatial graphs to evolve strategic ground water management for urban environments like Hyderabad city. We apply the statistical tools such as metamodelling and simulation to observe the variations in different areas and the impact of influencing factors.

3. Materials and Methods

3.1. Spatial graph theory

Spatial graph provides a graph in the 3-dimensional Euclidean space R^3 or the 3-sphere. For a graph G we take an embedding $f : G \to R^3$ then the image $\overline{G} = f(G)$ is called a spatial graph of G. It is a generalization of knot and link. Several parameters are influenced the quality of water index. Water quality can be assessed using spatial graph algorithms (Fall et al. 2007, Dale and Fortin, 2010) structural patches (nodes) and the Euclidean distances between them (links). We describe the qualitative spatial model. The entities and spatial relations represent the nodes and edges of the graph,

respectively. We describe the entities and various spatial relations included in the model at different graphical representations of the water quality. Measures of nodes which have the high patches indicates that the quality of water index. Water quality can be assessing excellent, good, poor, very poor, and unfit for human consumption of water quality in the selected area and which can be indicated with different nodes (shown in Figure 2).

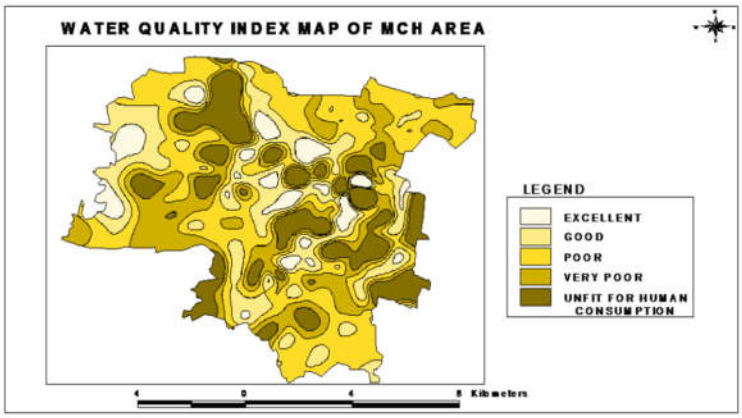

Figure 2 The water quality index of studied area

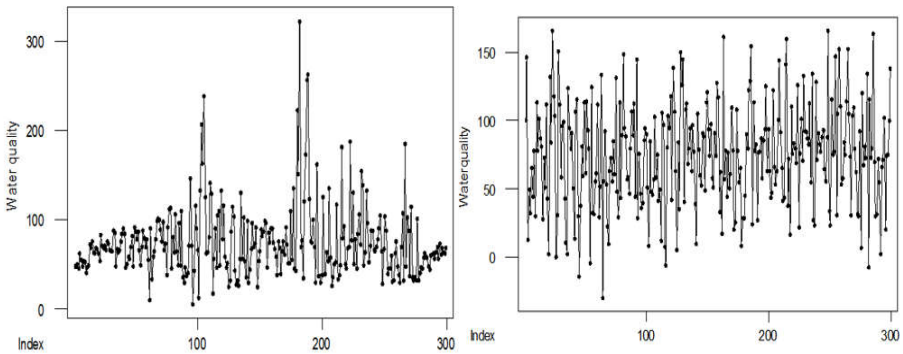

Figure 3 Actual and simulated pattern

By observing the spatial graph we identified that in the selected area most of the areas has shown unfit for human consumption. In the simulated graph, we clearly noted that the different factors such as CI, NO_3 and K shows high impact on water quality. In the usual regression, it has shown less variation but in the simulated graph the high impact of water quality by different parameters has clearly shown in Figure 3 (Actual and simulated data pattern) and also showed that the low, medium and high sodium concentration levels in the ground water.

3.2. Metamodelling

Suppose that there is a simulation output response variable, Y that is related to k independent variables say $x_1, x_2, ..., x_k$. The dependent variable Y, is a random variable, where the independent variables $x_1, x_2, ..., x_k$ are called design variables and are usually subject to control. The true relationship between the variables Y and x is represented by the simulation model. Our goal to

approximate this relationship by a simpler mathematical function called a metamodel. Regression analysis is one method for estimating the parameters.

The function relationship of the for m of several independent variables influence the response variable is

$$Y = b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_i x_i, \text{ where } i = 1, 2, \ldots, m. \quad (1)$$

In the present paper, we consider the water quality as response variable (Y), which is influence by the different variables such as the concentrations of PH (x_1), Electrical conductivity (EC) (x_2), Alkalinity (x_3), Chloride (x_4), NO_3 (x_5), Total dissolved solids (x_6), Hardness (x_7), Na^+ (x_8), K^+ (x_9), SO_4^{2-} (x_{10}), F^- (x_{11}) are as regressors. These all factors are affecting ground water quality and its index.

3.3. Random number assignment for regression

The assignment of random-number seeds or streams is part of the design of simulation experiment. Assigning a different stream to different design points guarantees that the responses *Y* from different design points will be statistically independent. Similarly, assigning the same stream to different design points induces dependence among the corresponding responses.

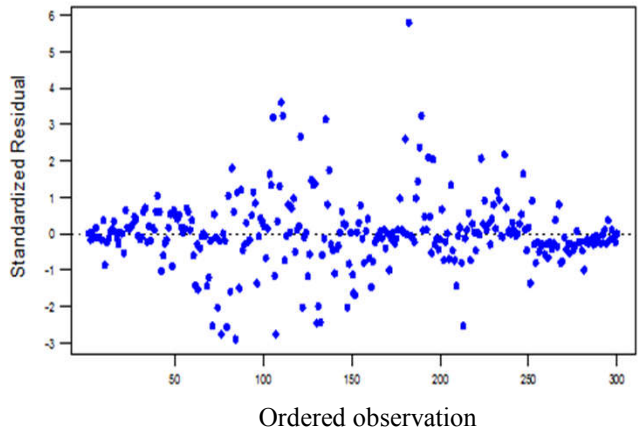

Figure 4 Water quality response and its residuals

3.4. Statistical analysis

The function relationship of the form of several independent variables influence the response variable is

$$Y = b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_i x_i, \text{ where } i = 1, 2, \ldots, 11. \quad (2)$$

The true or actual fitted values regression equation is

$$\hat{Y} = 11.2 + 1.69x_1 - 0.0017x_2 + 0.0085x_3 + 0.0076x_4 + 0.0253x_5 - 0.0017x_6 + 0.0041x_7 + 0.0412x_8 + 0.189x_9 - 0.0672x_{10} + 28.3x_{11}. \quad (3)$$

The fitted regression equation for the simulated data is

$$\hat{Y} = 193 - 14x_1 - 0.0111x_2 - 0.0083x_3 - 0.00766x_4 + 0.0205x_5 - 0.0042x_6 - 0.0162x_7 + 0.0995x_8 + 0.007x_9 - 0.0194x_{10} - 0.73x_{11}. \quad (4)$$

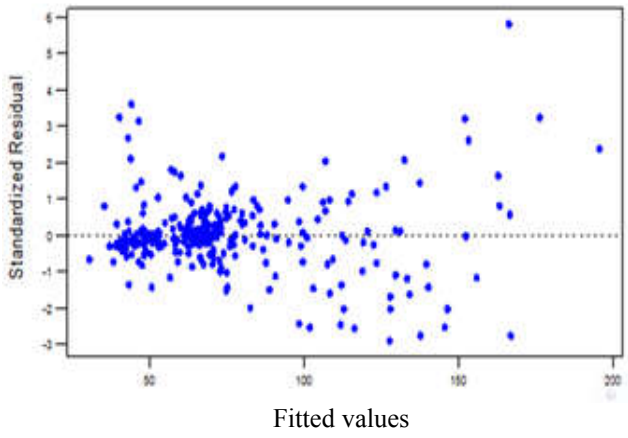

Figure 5 Residuals versus fitted values

The above empirical relation can be obtained by using statistical software Minitab 17 (Minitab 17 Statistical Software 2010). The following Table 1 representing the variations among the actual and simulated data.

Table 2 shows that the analysis of variance for the influenced parameters of actual and simulated data. As per the obtained results we conclude that the parameters of actual data showing that there is a significant difference between the several parameters which are affecting the water quality index. After simulating the data we got the results that there is no significant difference between the several parameters. Hence, we conclude the results on the basis of ANOVA that the simulation will be more powerful technique in reducing the significance difference among the groups has shown in the following table.

Table 1 Variations among actual and simulated data

Variable	N	Actual data		Simulated data	
		Mean	Standard deviation	Mean	Standard deviation
X_1	300	7.23	0.40	7.26	0.41
X_2	300	1137.10	483.30	1125.40	485.50
X_3	300	282.73	112.99	277.52	110.61
X_4	300	155.40	258.70	142.10	270.40
X_5	300	84.01	85.83	88.18	86.54
X_6	300	720.00	258.30	723.50	264.80
X_7	300	396.50	154.75	403.18	162.50
X_8	300	75.07	42.86	73.08	43.76
X_9	300	10.92	13.68	11.25	13.68
X_{10}	300	39.41	29.76	40.24	28.98
X_{11}	300	1.54	1.04	1.60	1.10

Table 2 ANOVA table

	Actual data				Simulated data		
Source	DF	SS	MSE	F	SS	MSE	F
Regression	11	272,892	24,808	32.36	29,776	2,707	1.87
Error	288	220,820	767		417,089	1,448	
Total	299	493,712			446,866		

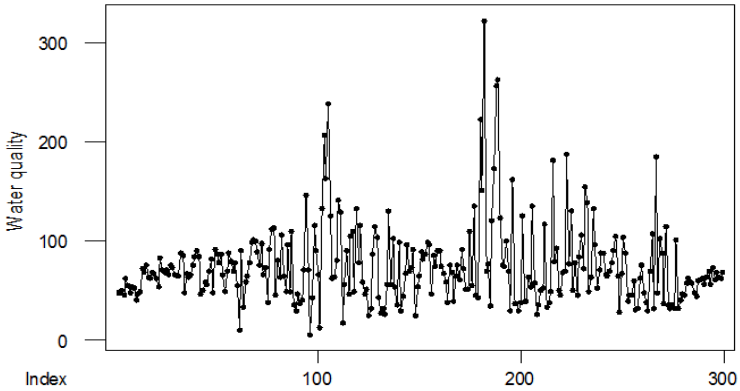

Figure 6 Actual data

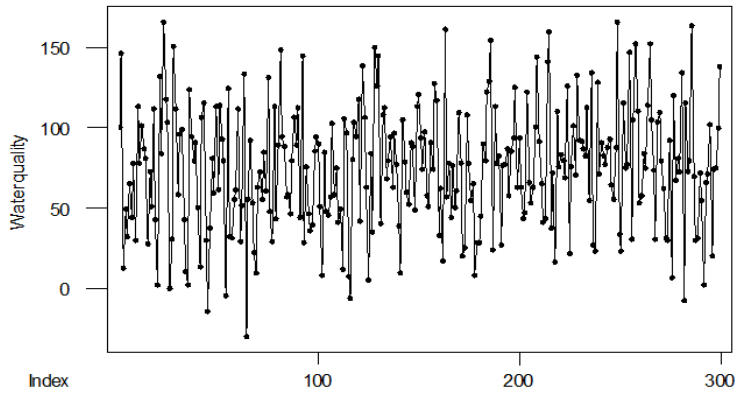

Figure 7 Simulated data

From the above calculations, we observed that X_4 (CI), X_5 (NO_3) and X_9 (k) has large variations which are more effecting the water quality index. By considering the mean value of each parameter and simulating the data and finding mean and variance of the water quality index, which are determined by mean = 148.3181 and standard deviation = 41.67. Out of 300 areas 25 observations showed the large standardized residual and 8 observations have larger influence in effecting the water quality and identified that most of the influenced areas in the collected or true data are unfit for consumption. After simulating the data, out of 300 areas only 14 areas showing the large standardized residual and one area have larger influence in effecting the water quality. In Table 3, the index of water quality has shown in the actual data is good 78, excellent 5, poor 101, unfit for human consumption 51, very poor 65, similarly for the simulated data is good 55, excellent 30, poor 65, unfit for human consumption 76, very poor 74.

Table 3 Quality index of actual and simulated data

Quality index	Probability	
	Actual data	Simulated data
Good	0.26	0.18
Excellent	0.02	0.10
Poor	0.34	0.22
Very poor	0.21	0.24
Unfit for Consumption	0.17	0.25
Actual data: S = 27.69, R^2 = 55.3% and R^2 (adj) = 53.6%		
Simulated data S = 38.06, R^2 = 6.7% and R^2 (adj) = 3.1%		

Table 4 Correlation matrix

	PH	EC	ALKA LINITY	CI	NO_3	TDS	HARD NESS	NA^+	K^+	SO_4^2
EC	−0.33									
ALKALINITY	−0.21	0.56								
CIO	−0.02	0.24	0.17							
NO_3	−0.20	0.46	0.12	0.05						
TDS	−0.35	0.87	0.67	0.25	0.46					
HARDNESS	−0.33	0.69	0.55	0.20	0.44	0.73				
NA^+	−0.30	0.63	0.51	0.23	0.33	0.66	0.29			
K^+	−0.08	0.29	0.30	0.04	0.18	0.33	0.17	0.24		
SO_4^2	−0.22	0.60	0.33	0.20	0.40	0.53	0.50	0.35	0.25	
F	−0.02	0.29	0.40	0.05	0.05	0.36	0.23	0.35	0.03	0.21

4. Conclusions

In the present paper, we emphasized on the groundwater quality, sources of ground water contamination, variation of groundwater quality and its spatial distribution, various effects of poor quality of groundwater, impacts of land use/land cover changes on quality of ground water. The area under study is highly urbanized, industrialized and fast developing city of Hyderabad and Secunderabad from the above calculations, we observed that X_4 (CI), X_5 (NO_3) and X_9 (K) has large variations which are more effecting the water quality index. Out of 300 areas 26 observations showed the large standardized residuals and 8 observations have larger influence in effecting the water quality and identified that most of the influenced areas in the collected or true data are unfit for consumption. After simulating the data, out of 300 areas only 14 areas showing the large standardized residuals and one area have larger influence in effecting the water quality. The index of water quality has shown in Table 3 and the results indicating that the actual data is good 26%, excellent 2%, poor 34%, unfit for human consumption 17%, very poor 21%. Similarly, for the simulated data is good 18%, excellent 10%, poor 22%, unfit for human consumption 25%, very poor 24%. Here we observed that the large residuals can be removed by applying the simulation and getting the accurate result. Correlation matrix depicts the relationship of each every factor and their relations which has provided in Table 4.

Hence we conclude that the spatial graph theory, regression and simulation will help us to give the accurate result when more variables are affecting the water quality index and it is possible to identify the areas which have low quality of water index and unfit for human consumption. It will be helpful to take the precautionary steps for better use of resources. We identified minimum water quality index with 4.55 identified at Erramanzil it is representing the area has excellent quality of water and maximum showed 323 at Bathkammakunta in the area of study region. Here we used spatial graphs to analyze the water quality parameters and its impact on quality index. Spatial graphs have been obtained by using remote sensing and GIS. From the above analysis we can conclude that, out of

300 areas 26 observations showed the large standardized residuals and 8 observations have larger influence in effecting the water quality index and also observed that possible interactions are with variable NA, K, F and overall lack of fit test is significant. R^2 is quite impressive and residuals appear to exhibit the cyclical pattern about the regression line. The above all results are obtained based on the range we mentioned in the following Table 5. It clearly depicts the areas located in the specified zones like excellent, good, poor, very poor and unfit for drinking. We observed from Table 5 that the areas which have excellent water quality index among our study region and also noted that the areas with specific quality index.

Table 5 The following table represents the areas which have different range of water quality indices

0-25	Excellent	Bodiwadi, Motinagar Begumpet, Erramanzil and Mehdipatnma
25 – 50	Good	Afzalgunj, Abids,Bahadurpura, Mirchowk, Khairtabad, Puranapul, Kachiguda, Secretariat etc.
50 – 75	Poor	Humayunnagar, Ramdevguda, Sanathnagar, Nayapul, Kamalapuri, Shaikpet, Chandrayangutta, Erragadda etc.
75- 100	Very Poor	Chintalbasti, Azamabad, Lalaguda, Ashoknagar Musheerabad, Saidabad, Maredpally, Langar House, Adikmet etc.
>100	Unfit for Drinking	Falaknuma, Chikadpally, Baghlingampally Zamistanpur, Gaganmahal, Jimkalvada, Golnaka etc.

Acknowledgments

The authors would like to thank the reviewers for their valuable comments and suggestions to improve the quality of the paper and presentation.

References

Anselin Luc, Ibnu S, Youngihn K. GeoDa: an introduction to spatial data analysis, Geog Anal. 2006; 38(1): 5-22.

Asadi SS, Padmaja V, Anjireddy M. Water quality index based assessment of ground water quality in municipal corporation of Hyderabad, India. J Ind Assoc Envir Mana. 2005; 32: 166-170.

Dormann CF, Mcpherson JM, Araujo MB et al. Methods to account for spatial autocorrelation in the analysis of species distributional data: a review. Ecog. 2007; 30: 609-628.

Erica F, Thomas W, Mattman BM et al. Recent developments in spatial graph theory, Cont Math. 2017; 689: 81-102.

Fall A, Fortin MJ, Manseau M, Obrien D. Spatial graphs: principles and application for habitat connectivity. Ecos. 2007; 10: 448-461.

Gyananath G, Islam SR, Shewdikar SV. Assessment of environmental parameter on ground water quality. Indi J Envi Prot. 2001; 21(4): 289-294.

Jerry B, Carson JS, Barry LN et al. Discrete-event System simulation. Singapore: Pearson Prentice Hal; 2010.

Kamaldeep SO, Geraldine DM, Yohan D, Pascal V. Spatial modeling of urban road traffic using graph theory. SAGEO-Rouen. 2017; 6-9.

Kobayashi K. On the spatial graph. Kodai Math J. 1994; 17: 511-517.

Kleijnen, JPC. Design and analysis of simulation experiments. New York: Springer; 2015.

Marie-Josee F, Patrick MAJ, Alistair M et al. Spatial statistics, spatial regression, and graph theory in ecology. Spa Stat. 2012; 1: 100-109.

Minitab 17 Statistical Software [Computer software]. State College, PA: Minitab, Inc. (www.minitab.com), 2010.

Russell RB. Tutorial: simulation metamodelling. Proceedings of the Winter Simulation Conference. USA: IEEE; 2015. 1765-1769.

Thomas ML, Ralph WK. Remote sensing and image interpretation. New York: John Wiley, 2000.