Performance of a negative binomial-GLM in spatial scan statistic: a case study of low-birth weights in Pakistan

Sami Ullah,[1] Mushtaq Ahmad Khan Barakzai,[2] Tianfa Xie[1]

[1]School of Mathematics, Statistics and Mechanics, Beijing University of Technology, China; [2]Department of Mathematics and Statistics, Institute of Business Management, Karachi, Pakistan

Abstract

Spatial cluster analyses of health events are useful for enabling targeted interventions. Spatial scan statistic is the state-of-the-art method for this kind of analysis and the Poisson Generalized Linear Model (GLM) approach to the spatial scan statistic can be used for count data for spatial cluster detection with covariate adjustment. However, its use for modelling is limited due to data over-dispersion. A Generalized Linear Mixed Model (GLMM) has recently been proposed for modelling this kind of over-dispersion by incorporating random effects to model area-specific intrinsic variation not explained by other covariates in the model. However, these random effects may exhibit a geographical correlation, which may lead to a potential spatial cluster being undetected. To handle the over-dispersion in the count data, this study aimed to evaluate the performance of a negative binomial-GLM in spatial scan statistic on real-world data of low birth weights in Khyber-Pakhtunkhwa Province, Pakistan, 2019. The results were compared with the Poisson-GLM and GLMM, showing that the negative binomial-GLM is an ideal choice for spatial scan statistic in the presence of over-dispersed data. With a covariate (maternal anaemia) adjustment, the negative binomial-GLM-based spatial scan statistic detected one significant cluster covering Dir lower district. Without the covariate adjustment, it detected two clusters, each covering one district. The district of Peshawar was seen as the most likely cluster and Battagram as the secondary cluster. However, none of the clusters were detected by GLMM spatial scan statistic, which might be due to the spatial correlation of the random effects in GLMM.

Correspondence: Sami Ullah, School of Mathematics, Statistics and Mechanics, Beijing University of Technology, Beijing 100124, China.
Tel.: +86.17813072535.
E-mail: sami.khan3891@gmail.com

Key words: spatial scan statistics; low birth weight; negative-binomial-GLM; Pakistan.

Conflict of interest: the authors declare no potential conflict of interest, and all authors confirm accuracy.

Availability of data and materials: all data generated or analyzed during this study are publically available at: https://www.dgh-skp.gov.pk/reports.html

Funding: this work was funded by the National Social Science Fund of China under grant No. 21BTJ041.

Received: 12 May 2024.
Accepted: 8 August 2024.

©Copyright: the Author(s), 2024
Licensee PAGEPress, Italy
Geospatial Health 2024; 19:1313
doi:10.4081/gh.2024.1313

This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License (CC BY-NC 4.0).

Publisher's note: all claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article or claim that may be made by its manufacturer is not guaranteed or endorsed by the publisher.

Non-commercial use only

Introduction

According to the World Health Organization (WHO), birth weights less than 2500 g are considered Low Birth Weights (LBWs). Out of an estimated 130-140 million annual births globally, 20 million or 15-20% of new born children are considered LBWs, with the highest proportion in low-income countries (Blencowe *et al.*, 2019; WHO, 2014). Pakistan Demographic and Health Survey (NIPS) found a high proportion of the global LBW burden, with 19% in urban and 32% in rural areas (NIPS, 2013). The high LBW rate in Pakistan has shown adverse health consequences and impeded efforts to achieve United Nation's Millennium Development Goals (MDGs). It is also, likely to affect Pakistan's ability to achieve the health and nutrition part of the updated Sustainable Development Goals (SDGs) (Planning Commission, 2013).

Some studies on the risk factors of LBW have identified maternal anaemia as one of the risk factors of LBW (Baig *et al.*, 2020; Iqbal *et al.*, 2023). Anaemic women had a 6.8-fold increased incidence of LBW as reported by a Nepalese study (Rana *et al.* 2013). Another study revealed that anaemia during pregnancy raises the newborn's risk of LBW (Engidaw *et al.*,2022). In Pakistan, several studies have been conducted to investigate these associated risk factors (Rashid *et al.*, 2020; Iqbal *et al.*, 2022; Zahra *et al.*, 2022). However, none of the studies have focused on the geographical cluster analysis of LBW in Pakistan. Using maternal aneemia as a covariate, our aim was to detect the covariate-adjusted geographical clusters of LBW in Khyber-Pakhtunkhwa Province, Pakistanin 2019.

For this kind of analysis, many statistical techniques have been proposed (Grimson 1993; Kulldorff, 1997; Anderson *et al.*, 1997; Cook *et al.,* 2007; Huang *et al.*, 2007; Jung *et al.*, 2007). A common issue with these techniques is covariate adjustment, *e.g.*, non-randomly distributed covariates associated with a disease, or other public health event, may lead to the detection of inaccurate spatial clusters by the spatial distribution of these covariates. Spatial scan statistic (Kulldorff, 1997) is one of the most popular techniques for spatial cluster analysis (Ishioka *et al*., 2019; Li *et al*., 2019; Leyso *et al*., 2020; Frévent *et al*., 2021; Ullah *et al*., 2020, 2021). It is available for Bernoulli, Poisson, normal, exponential and ordinal models. In some scenarios, covariate adjustment is possible in Spatial Scan Statistic (Kulldorff, 1997; Klassen *et al*., 2005; Sheehan *et al*., 2005). A Generalized Linear Model (GLM) has been proposed for spatial scan statistic approaches to adjust for covariates using different probability models such as Poisson, Bernoulli, normal and gamma (Jung 2009; Zhang *et al.*, 2009). In this approach, the test statistic for a cluster is equal to fitting a GLM using a cluster variable as a predictor. This cluster variable is a dummy variable that has a value of 1 for regions inside the cluster and 0 for regions outside the cluster. The inclusion of cluster covariates in the model yields an estimate of the elevated risk as measured by the corresponding coefficient, and its statistical significance as measured by the corresponding *p*-value.

The GLM-Poisson-based spatial scan statistic is widely used for count data. However, the GLM-Poisson model has limitations for modelling over-dispersion in the count data and hence may not be an ideal choice for the spatial scan statistic in cases of of over-dispersion. To handle this, a GLMM has recently been used (Gómez-Rubio *et al*., 2019). It incorporates random effects to model area-specific intrinsic variation not explained by other covariates or cluster variables in the model. However, these random effects may exhibit geographical correlations (Bilancia *et al*., 2014), which may lead to a potential spatial cluster being undetected.

Studies have shown that a Negative Binomial model can handle the over-dispersed counts by including an additional dispersion parameter (An *et al*., 2016; Stoklosa *et al.,* 2022). Because of the quadratic nature of the mean-variance relationship, the negative binomial model is a useful approach for modeling the over-dispersion in the count data. This study aims to evaluate the performance of a Negative Binomial-GLM in the Spatial scan statistic for detecting potential spatial clusters of LBWs at the district level in Khyber-Pakhtunkhwa Province, Pakistan in 2019.

Materials and Methods

Approach

Negative Binomial model was compared with Poisson GLM and GLMMs,with the analyses executed in R software.

Study site

The province of Khyber Pakhtunkhwa has a total area of 74,521 km^2 and a population of 30.52 million, according to the 2017 census. The land area of the province is traditionally divided into 25 districts. However, in 2019, the seven Federally Administered Territories (FATA) were politically and administratively merged into the province of Khyber Pakhtunkhwa.

Data

Data for 2019 were collected on LBW cases, maternal anemia and population-at-risk at the district level from an annual report of the District Health Information System (DHIS), Khyber Pakhtunkhwa Province, Pakistan (DHIS, 2019). Total births and pregnant women in each district were used as the population-at-risk for LBW and maternal anaemia, respectively. The DHIS collects monthly data on reported LBW cases and other diseases from all government hospitals in each district, which are eventually filed in the respective provincial office. The percentage data on LBW and maternal anaemia in each district of the province during the year 2019 are shown on the chart in Figure 1.

Spatial scan statistics

In this study, we used a GLM approach to Spatial Scan Statistic (Gómez-Rubio *et al*., 2019; Jung *et al*., 2009) for detecting the spatial clusters of LBW using maternal anaemia as a covariate. Spatial Scan Statistic scans the study area through a moving circular window. The circular scanning window was centred on each subregion centroid with the radius varying in size from zero up to some fixed value, which resulted in a large number of overlapping circular zones of different sizes. The regions with centroids within the circular window were considered as belonging to that circular zone. Each circular zone suggests a potential candidate cluster. Given a possible candidate cluster, say *z,* the spatial scan statistic compares the relative risk within the candidate cluster θ_z and outside the candidate cluster , based on the test given below:

$$H_0 : \theta z = \theta_{\bar{z}}$$

$$H_1 : \theta z > \theta_{\bar{z}}$$

This test is performed via a likelihood ratio statistic, where many different possible clusters are tested in turn by changing the areas in *z*, and the most likely cluster (*i.e.* the one with the highest value of the test statistic) was selected. The significance of this cluster was assessed with a Monte Carlo test that also provides the *p*-value.

Spatial scan statistic uses different models such as binomial,

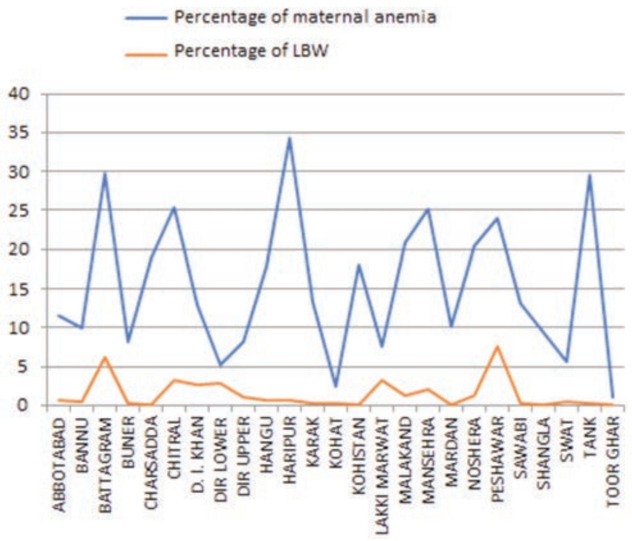

Figure 1. The percentages of LBW and maternal anemia over the districts in Khyber-Pakhtunkhwa Province, Pakistan.

Non-commercial use only

Poisson, Gaussian or negative models for calculating the likelihood ratios for each circular zone. Several studies provided an explicit link between GLMs and the spatial scan statistic (Jung *et al*., 2009; Zhang *et al*., 2009; Gómez-Rubio *et al.,* 2019). GLMs not only model Poisson or binomial responses, but can also link the outcome to a linear predictor on the covariates (and, possibly, other effects). The spatial scan statistic provides different results for different models. We implemented it with a negative binomial-GLM.

Negative binomial-GLM

The details of the negative binomial regression model have been described by Nava *et al*. (2014). A negative binomial-GLM is commonly used to analyze count data, particularly when the data exhibit over-dispersion, meaning the variance is larger than the mean. This is often the case with disease counts, where the number of occurrences can vary widely across different regions or periods. In the negative binomial regression model, we modelled the log of the expected count (μ_i) as a linear combination of the predictors and incorporated the exposure term as an offset representing the population-at-risk. Additionally, this model introduces an additional parameter θ to capture the extra variability (over-dispersion) in the count data as expressed by Equation 1.

$$\log(\mu_\mathrm{i}) + \log(\mathrm{E_i}) + \alpha + \beta_\mathrm{i}\mathrm{x_i} + \theta \qquad \text{Eq. 1}$$

where μ_i denotes the expected value of the response variable (in this case, the count of LBW cases) for the i^{th} region; E_i the exposure or offset variable, representing the underlying population size for the i^{th} region; α the intercept term; β_i the coefficient associated with the predictor variable x_i; θ an additional parameter to account for over-dispersion.

Results and Discussion

Model fitting

First, we fitted a Poisson-GLM to the LBW data and presented the results (Table 1). Then the data were tested for over-dispersion using the P_B score test as proposed by Dean *et al.* (1992). The results showed a test statistic value, *P_B* = 528.32 with a *p*>0.001, indicating evidence of over-dispersion. These results lead us to use a negative binomial approach to model this type of over-dispersed data. In order to choose an appropriate model for the spatial scan statistic of LBW cases in the scenario of over-dispersion, we fitted the negative binomial-GLM and GLMM-Poisson to the LBW data. The comparison of the results is shown in Table 2. The results show that both models provide approximately similar values for the coefficient's estimates, Standard error (SE), as well as Akaike's Information Criterion (AIC) values, indicated that the negative binomial-GLM can be effective for modelling the over-dispersion in our data. It should be noted that the GLMM captures the over-dispersion by incorporating the random effects in the model, which may exhibit geographical correlation and hence may lead to the potential spatial clusters being undetected. Moreover, the covariate maternal anaemia showed a significant impact on the LBW cases at the 5% level of significance and hence can be a useful covariate for modelling the LBW data. The positive estimate for this covariate shows that LBW cases increase with the increase in maternal anaemia.

Spatial cluster detection with covariate adjustment

In this section, we applied the negative binomial-GLM and GLMM in the spatial scan statistic to detect the potential spatial clusters of LBW cases adjusted for a covariate (maternal anaemia) in Khyber Pakhtunkhwa Province in 2019. The results are shown in Table 3. These results show that negative binomial-GLM detected the Dir lower district as a spatial cluster of LBW cases with a risk of 1.43. However, the GLMM failed to detect any clusters, possibly because of the spatial correlation of random effects. These results show that when disease counts are over-dispersed relative to the Poisson distribution, a negative binomial-GLM is the ideal model to be used in spatial scan statistic.

Spatial cluster detection without covariate adjustment

In order to know the influence of using the covariate (maternal anaemia) in spatial cluster detection of LBW cases, we applied the spatial scan statistic with the negative binomial-GLM without a covariate adjustment to LBW data. Without this covariate, two

Non-commercial use only

Table 1. The outcomes of the Poisson GLM on LBW data.

Model	Coefficient	Estimate	SE	z-value	*p*-value	AIC
Poisson-GLM	Intercept	-7.43e^{-01}	2.95e-0^{2}	-25.18	<0.001	3337.8
	Maternal anemia	1.16e-04	3.05e-^{06}	38.13	<0.001	

GLM, Generalized Linear Model; SE, standard error.

Table 2. Comparison of a negative binomial-GLM and GLMM-Poisson on LBW data

Model	Coefficient	Estimate	SE	Test-statistic	*p*-value	AIC
Negative Binomial- GLM	Intercept	-8.1e^{-01}	3.77e^{-01}	-2.15	0.040	276.45
	Maternal anaemia	1.34e^{-04}	5.80e^{-05}	2.27	0.030	
GLMM	Intercept	-1.57e^{00}	3.95e^{-01}	-3.97	<0.01	276.7
	Maternal anaemia	1.48e^{-04}	5.97e^{-05}	2.48	0.013	

LBW, low birth weight; GLM, Generalized Linear model; GLMM, Generalized Linear Mixed Model; SE, standard error.

clusters were detected, each covering one district as given in Table 4. The district of Peshawar was seen as the most likely cluster with a risk of 1.65 and Battagram as a secondary cluster with a risk of 1.50. These two districts exhibited a high percentage of maternal anaemia as shown in Figure 1. Therefore, after adjusting for maternal anaemia, these districts were not detected as the potential spatial clusters of LBW cases. Previous studies have also reported these districts as potential clusters of maternal anaemia (Ullah *et al*., 2023). It suggests that these districts were seen in our analysis as spatial clusters of LBW due to the presence of maternal anaemia. In addition, these results indicate the importance of the covariate (maternal anaemia) in the spatial cluster analysis of LBW in the study area. The locations of the most likely cluster are shown in Figure 2.

The spatial cluster analyses with a covariate adjustment detected a new cluster that was not apparent in the analysis without a covariate adjustment (see in Tables 3 and 4). These results show that the LBW clusters detected without being adjusted for maternal anaemia might lose their significance when maternal anaemia is accounted for, indicating that this is a major factor of LBW cases in those areas. Hence, cluster detection of LBW cases without covariate adjustment helps identify general areas of concern, while adjusting for maternal anaemia isolates the impact of maternal anaemia, allowing for more precise public health interventions.

Conclusions

This study proposes the implementation of spatial scan statistic with the negative binomial-GLM in scenarios characterized by data over-dispersion. The GLMM can capture the over-dispersion by including the random effects in the model; however, it may not be able to detect a potential spatial cluster if these random effects are spatially correlated. Moreover, maternal anaemia was found to be a useful covariate for spatial cluster analyses of LBW data. Without a covariate adjustment, the districts of Peshawar and Battagram were seen as potential LBW spatial clusters which might be due to the high volume of maternal anaemia. However, adjusting for this covariate, the district of Dir lower was seen as the purely spatial cluster of LBW. Without the covariate adjustment, completely different results were obtained showing the importance of maternal anaemia in the spatial cluster analysis of LBW cases. This study suggests Dir lower district to be the most suitable region for possible interventions to control the LBW cases in the province. In addition, it provides important insights for policymakers to address the health issues related to LBW in the targeted region. Investigation of other covariates of LBW is recommended for future work as it can help in more accurate LBW cluster analysis.

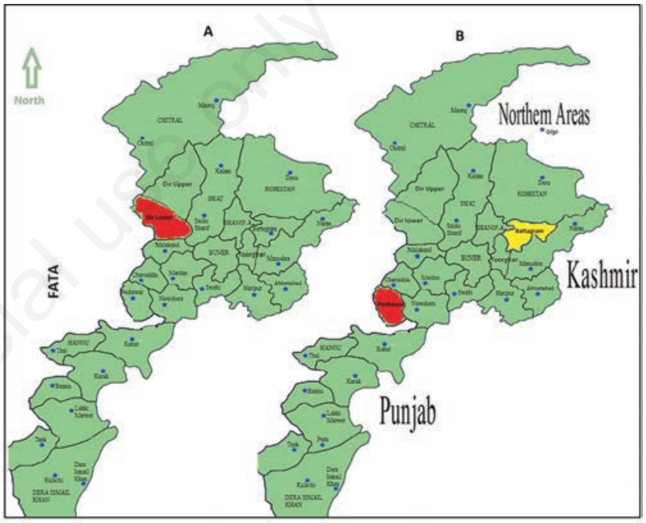

Figure 2. The geographical locations of LBW clusters. A) with a covariate adjustment; B) without a covariate adjustment; Red clour indicates the most likely cluster; Yellow indicates the secondary cluster.

Non-commercial use only

Table 3. Spatial clusters of LBW with covariate adjustment.

Model	Cluster no.	District	Statistic	*p*-value	Risk
Negative Binomial-GLM	1	Dir Lower	1.409	0.09	1.427
GLMM	-	-	-		

GLM, Generalized Linear model; GLMM, Generalized Linear Mixed Model.

Table 4. Spatial clusters of LBW without covariate adjustment.

Cluster no.	District	Statistic	*p*-value	Risk
1	Peshawar	1.80	0.04	1.657
2	Battagram	1.30	0.09	1.50

LBW, low birth weight.

References

Anderson NH, Titterington DM, 1997. Some methods for investigating spatial clustering, with epidemiological applications. J R Stat Soc Ser A: Stat Soc 160:87-105.

An Q, Wu J, Fan X, Pan L, Sun W, 2016. Using a negative binomial regression model for early warning at the start of a hand foot mouth disease epidemic in Dalian, Liaoning Province, China. PLoS One 11: e0157815.

Baig JA, Jamal MM, Jamal J, Musarrat M, 2020. To determine the association of maternal anemia with perinatal outcome in tertiary care hospital. Pak. Armed Forces Med. J 70:302-7.

Bilancia M, Demarinis G, 2014. Bayesian scanning of spatial disease rates with Integrated Nested Laplace Approximation (INLA). Stat Methods Appt 23:71-94.

Blencowe H, Krasevec J, De Onis M, Black RE, An X, Stevens GA, Borghi E, Hayashi C, Estevez D, Cegolon L, Shiekh S, 2019. National, Regional, and Worldwide Estimates of Low Birthweight in 2015, with Trends from 2000: A Systematic Analysis. Lancet Glob Health 7:e849-60.

Cook AJ, Gold DR, Li Y, 2007. Spatial cluster detection for censored outcome data. Biometrics 63:540-49.

Dean CB, 1992. Testing for overdispersion in poisson and binomial regression models. J Am Stat Assoc 87:451-57.

DHIS KP, 2019. Annual Report. Available from: https://www.dghskp.gov.pk/reports.html

Engidaw MT, Eyayu T, Tiruneh T, 2022. The effect of maternal anaemia on low birth weight among newborns in Northwest Ethiopia. Sci Rep 12:15280.

Frévent C, Ahmed MS, Marbac M, Genin M, 2021. Detecting spatial clusters in functional data: new scan statistic approaches. Spat Stat 46:100550.

Gómez-Rubio V, Moraga P, Molitor J, Rowlingson B. 2019. DClusterm: model-based detection of disease clusters. J Stat Softw 90:1-26.

Grimson RC, 1993. Disease clusters, maxima, and p-values. Stat Med 12:1773-94.

Huang L, Kulldorff M, Gregorio D, 2007. A spatial scan statistic for survival data. Biometrics 63:109-18.

Iqbal A, Kanwal M, Rani N, Abbas S, Lakhan S, Pawan N. A, 2023. A prospective study to assess the correlation of neonatal birth weight with the hemoglobin level of the mother during pregnancy. Pakistan J Med Health Sci 17:799.

Iqbal S, Tanveer A, Khan Z, Junaid KM, Mushtaq N, Ali N, 2022. Risk factors of low birth weight in Pakistan. Pakistan J Med Health Sci 16:1163.

Ishioka F, Kawahara J, Mizuta M, Minato SI, Kurihara K, 2019. Evaluation of hotspot cluster detection using spatial scan statistic based on exact counting. Jpn J Stat Data Sci 2:241-62.

Jung I. 2009. A generalized linear models approach to spatial scan statistics for covariate adjustment. Stat Med 28:1131-43.

Jung I, Kulldorff M, Klassen AC, 2007. A spatial scan statistic for ordinal data. Stat Med 26:1594-1607.

Klassen AC, Kulldorff M, Curriero F, 2005. Geographical clustering of prostate cancer grade and stage at diagnosis, before and after adjustment for risk factors. Int J Health Geogr 4:1-16.

Kulldorff M, 1997. A spatial scan statistic. Commun Stat Theory Methods 26:1481-96.

Leyso NLC, Palatino MC, 2020. Detecting local clusters of under-5 malnutrition in the province of Marinduque, Philippines using spatial scan statistic. Nutr Metab Insights 13:1178 638820940670.

Li M, Shi X, Li X, Ma W, He J, Liu T, 2019. Sensitivity of disease cluster detection to spatial scales: an analysis with the spatial scan statistic method. Int J Geogr Inf Sci 33:2125-52.

Nava M. 2014. Generalized Linear Regression Models for Count Data. California State University, Long Beach; 2014. Available from: https://books.google.com/books?id=DaB90AEACAAJ

NIPS D, 2013. Pakistan Demographic and Health Survey 2012-13. Available from: https://dhsprogram.com/pubs/pdf/FR290/FR290.pdf

Planning Commission, MP, 2013. Pakistan Millennium Development Goals Report 2013. Available from: https://www.undp.org/pakistan/publications/pakistan-mdgs-report-2013

Rana SS, Sharma S, Chand A, Malla R, 2013. Relationship between maternal haemoglobin and fetal weight. Nepal Journal of Obstetrics &Gynaecology 8:37-40.

Rashid HU, Khan MN, Imtiaz A, Ullah N, Dherani M, Rahman A, 2020. Post-traumatic stress disorder and association with low birth weight in displaced population following conflict in malakand division, pakistan: a case-control study. BMC Pregnancy and Childbirth 20:1-8.

Sheehan TJ, DeChello LM, 2005. A space-time analysis of the proportion of late stage breast cancer in Massachusetts, 1988 to 1997. Int J Health Geogr 4:1-9.

Stoklosa J, Blakey R V, Hui FKC, 2022. An overview of modern applications of negative binomial modelling in ecology and biodiversity. Diversity 14:320.

Ullah S, Burney SMA, Rasheed T, Burney S, Barakzia MAK, 2023. Space-time cluster analysis of anemia in pregnant women in the province of Khyber Pakhtunkhwa, Pakistan (2014-2020). Geospat Health 18:1192.

Ullah S, Daud H, Dass SC, Fanaee-T H, Kausarian H, Alamgir, 2020. space-time clustering characteristics of tuberculosis in Khyber Pakhtunkhwa Province, Pakistan, 2015-2019. Int J Environ Res Public Health 17:1413.

Ullah S, Nor NHM, Daud H, Zainuddin N, Gandapur MSJ, Ali I, Alamgir. 2021. Spatial cluster analysis of COVID-19 in Malaysia (Mar-Sep, 2020). Geospat Health 16:961.

WHO, 2014. Global Nutrition Targets 2025: Low Birth Weight Policy Brief. Available from: https://media.tghn.org/articles/WHO_NMH_NHD_14.5_eng.pdf

Zahra T, Mumtaz U, Riffat N, Mushtaq F, Cheema MH, Mahmud T, 2022. Factors associated with low birthweight among newborns delivered at term in a tertiary care hospital in Lahore. J. Fatima Jinnah Med. Univ16:20-26.

Zhang T, Lin G, 2009. Spatial scan statistics in loglinear models. Comput Stat Data Anal 53:2851-58.

Non commercial use only