Global Scientific JOURNALS
GSJ: Volume 11, Issue 9, September 2023, Online: ISSN 2320-9186
www.globalscientificjournal.com

SPATIAL MODELING AND MAPPING OF COUNT DATA WITH CASES OF UNDER-REPORTING: A CASE OF DIABETES IN KENYA

Mbeche Collins Kaba
Department of Mathematics and Physical Sciences
Maasai Mara University, Kenya.
Kabaacolloz11@gmail.com

Omondi Joseph Ouno
Department of Mathematics and Physical Sciences
Maasai Mara University, Kenya.

Okenye Justin Obwoge
Department of Mathematics
Egerton University, Kenya.

Abstract: Diabetes is a significant public health issue in developing countries, with an increasing burden on the healthcare system. However, accurate reporting of diabetes cases is often hindered by under-reporting, particularly in rural areas where access to healthcare is limited. When dealing with count data, both under-reported and over-reported cases are encountered. If it is assumed that the count data obtained from the field is always true, then modeling it with other count-data models will be erroneous. This study aimed to improve the existing Poisson-Binomial mixture model by factoring in covariates to make it suitable to estimate the number of under-reported diabetes cases in each county of Kenya and map the distribution of these cases. The covariates used in the model include the education level, poverty index, and access to healthcare in respective counties, making the probability of reporting vary from one county to another. The data was obtained from the Kenya Diabetes Management Information Centre and Kenya National Bureau of Statistics. The results revealed that at least each of the 47 counties had under-reported the diabetes data, with the probability of reporting ranging from 0.9002423 for Migori County and 0.7164098 for Mombasa County. Nairobi and Mombasa counties reported the highest underreporting rate with 16,708 and 11,784 cases, respectively underreported, while Lamu had 1269 underreported cases, the least in all the 47 counties. The Deviance Information Criterion (DIC) was used to compare the original model and the improved model, whereby the improved model was found to be efficient since it had a smaller DIC value. The computed actual cases of diabetes revealed that Nairobi and Lamu had 179,604 and 7,038, respectively, representing the highest and lowest diabetes county in Kenya. The resulting maps identified high-risk areas for under-reporting and the general distribution of diabetes in Kenya, valuable information for policymakers and public health practitioners to target resources towards improving diabetes prevention and management in Kenya.

Keywords: Spatial; Mapping; Deviance Information Criterion; Diabetes; Underreporting

1. Introduction

Different techniques are used in epidemiology to study the patterns of diseases among human populations, such as descriptive and diagnostic analyzes [14]. Spatial disease mapping is one of the common models that public health use in studying different diseases [10, 13]. With mapping techniques, epidemiologists and public health experts can detect the relationship between people and their environment [3, 15].

A study by [5] reveals that the spatial patterns of the disease in question can significantly impact the nature and type of data collected. Although under-reporting and over-reporting of cases occur in different scenarios, under-reporting of disease cases has been termed an impediment in determining the actual spatial patterns of a given disease. With advanced technologies, developed countries ensure that the number of reported cases is almost accurate [11, 15]. However, the situation in a majority of African countries is different. The study by [5] further explains that some aspects that affect most African countries, resulting in underreporting, include inadequate medical funds, low medical knowledge, poverty, and stigmatization. The locals may also need more confidence in the existing medical institutions, hence deliberately avoiding disclosing such important information. With the increased under-reporting of disease cases, some of the count models developed to model the cases seem inaccurate [12, 16].

Different models, such as Kriging and the Empirical Bayes, have been developed to investigate the presence of spatial property on count data [8, 9]. However, most models assume that the data in question is correctly reported, ignoring the issue of under-reporting [17]. The study by [2] revealed that some of the cases of diabetes are not accounted for, making it difficult to know the risks of the disease in different areas. There is, therefore, a need to adjust the under-reported cases of diabetes in the country and develop a map to show the relative risk of the disease in different counties. Under-reporting of diabetes cases is common, particularly in rural areas with limited access to healthcare. This study was meant to improve the Poisson-Binomial mixture model to make it

suitable for estimating the number of under-reported diabetes cases in each county of Kenya and creating a map of under-reported cases using available data on reported cases and relevant covariates.

A study by [1] investigated the spatial patterns and factors associated with under-reporting diabetes cases in Kenya, using data from the 2015 Kenya Stepwise Survey of Non-Communicable Diseases. The study identified under-reporting as a significant issue in Kenya and recommended using spatial modeling to improve estimates of diabetes burden.

Another study titled "Evaluation of the completeness of diabetes-related reporting systems in Kenya" was conducted by [4] and assessed the completeness and accuracy of diabetes-related reporting systems in Kenya. The researchers evaluated the existing reporting mechanisms and identified areas for improvement to enhance the quality and accuracy of diabetes data [6, 7]. The study pointed out that some of the cases of diabetes within the communities have not been adequately documented, making it hard to budget for the disease in the country.

Furthermore, the Kenya National Diabetes Strategy report [2] acknowledged the issue of under-reporting. It emphasized the need for a comprehensive surveillance system to capture accurate and representative data on diabetes prevalence and burden in the country [. According to the report, most of the counties in Kenya lacked accurate data on the number of people with diabetes. The report recommended that the counties develop better disease recording systems that will help present the actual data on certain diseases affecting respective counties.

2. Methods

The study used a Poisson-Binomial mixture model to estimate the number of under-reported diabetes cases in each county of Kenya and map the distribution of these cases. The Poisson-Binomial mixture model is a statistical method used to model count data with over- or under-dispersion, a common issue with health data. The model will combine the Poisson distribution and the Binomial distribution to account for the mean and variance of the count data. The model will use data on reported diabetes cases and relevant covariates such as education level, poverty rate, and access to healthcare to estimate the number of under-reported cases in each county. The covariates will be included in the model as fixed effects to control for potential confounding effects. Spatial autocorrelation will be accounted for using a spatial random effect.

Consider a case where y_i^* represents the total cases of an event in unit i. Let y_i^* have a Poisson distribution. The main issue experienced here is the problem of under-reporting. Given that the number of reported cases is represented by y, it is essential that it does not represent the actual value of the disease count at unit i. It, therefore, means that y_i represents a fraction of the reported cases y_i^* in the unit i. We will therefore have the following binomial distribution:

$$P(y_i|y_i^*,\lambda_i) \sim Bin\,(y_i^*,\lambda_i)$$

Two different approaches can be used to estimate the number of reported cases. The number of reported cases equals the actual value, whereby $y_i = y_i^*$. On the other hand, the number of reported cases can be expressed as $y_i = y_i^* - n$ whereby $n < y_i^*$.

Assuming that the observed disease count has a binomial distribution, the marginal number of reported cases (y_i) will be given by:

$$P(Y_i = y) = \sum_{y^* \geq y}^{\infty} \binom{y^*}{y} p^y (1-p)^{y^*-y} \frac{\lambda^{y^*} e^{-\lambda}}{y^*!} \quad (1)$$

where y is the number of reported cases (observed disease count), y* is the total number of cases (both reported and unreported),

Factoring out gives:

$$P(Y_i = y) = \frac{e^{-\lambda} p^y}{(1-p)^y} \sum_{y^* \geq y}^{\infty} \binom{y^*}{y} (1-p)^{y^*} \frac{\lambda^{y^*}}{y^*!} \quad (2)$$

$$\frac{e^{-\lambda} p^y}{(1-p)^y} \sum_{y^* \geq y}^{\infty} \frac{y^*!}{y!(y^*-y)!} (1-p)^{y^*} \frac{\lambda^{y^*}}{y^*!} \quad (3)$$

$$\frac{e^{-\lambda} p^y}{(1-p)^y} \sum_{y^* \geq y}^{\infty} \frac{1}{y!(y^*-y)!} (1-p)^{y^*} \lambda^{y^*} \quad (4)$$

Now adding y and subtracting y on the powers with y^* now gives:

$$= \frac{e^{-\lambda} p^y}{y!(1-p)^y} \sum_{y^* \geq y}^{n-1} \frac{1}{(y^*-y)!} (1-p)^{y^*-y+y} \lambda^{y^*-y+y} \quad (5)$$

Now let $y^* - y = k$

$$= \frac{e^{-\lambda} p^y}{y!(1-p)^y} \sum_{k=0}^{n-1} \frac{1}{k!} (1-p)^{k+y} \lambda^{k+y} \quad (6)$$

$$= \frac{e^{-\lambda} p^y}{y!(1-p)^y} (1-p)^y \lambda^y \sum_{k=0}^{n-1} \frac{1}{k!} (1-p)^k \lambda^k \quad (7)$$

Factoring out gives:

$$\frac{e^{-\lambda}(p\lambda)^y}{y!} e^{\lambda(1-p)} = \frac{e^{-\lambda}(p\lambda)^y}{y!} e^{\lambda} e^{\lambda p} \quad (8)$$

This gives us a likelihood function for estimating underreporting which is given by:

$$\frac{(\lambda p)^y}{y!} e^{\lambda p} \quad where\ y \geq 0 \qquad (9)$$

Therefore, the observed counts (y) follow a Poisson distribution with mean as shown in equation (9). From the equation, ignoring under-dispersion, the count data on the number of diabetes cases in Kenya will be modeled using the Poisson model. Considering the under-dispersion, it varies spatially using the probability P where $P_i \neq P_j$.

Including covariates in the mixture, model allows for a more comprehensive and accurate estimation of the under-reporting of diabetes cases. By incorporating covariates, we can account for the potential effects of factors that influence the reporting behavior and vary across counties. This enables us to adjust the reporting probability based on the specific characteristics of each county, leading to more precise estimates of under-reporting. The significant covariates to be included in the model which might influence the rate of reporting of diabetes at the counties include education level (X1), poverty rate (X2), and access to healthcare (X3).

To specify the relationship between the covariates and the probability of reporting a diabetes case, logistic regression is used as follows:

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 \qquad (10)$$

The coefficients $\beta_0, \beta_1, \beta_2, and\ \beta_3$ capture the effects of the respective covariates on the probability of reporting a case in a given county.

By incorporating education level (X1), poverty rate (X2), and access to healthcare (X3) as covariates in our model, we aim to account for the influence of these factors on the reporting probability. This approach allows us to adjust the reporting probability based on the specific characteristics of each county, providing a more accurate estimation of under-reporting and a deeper understanding of the impact of education, poverty, and healthcare access on diabetes reporting patterns.

For λ, a gamma prior distribution will be used. The gamma prior for λ can be specified using two parameters, α and β. α determines the shape of the distribution, while β controls the rate of decay. By choosing appropriate values for α and β, you can reflect your prior beliefs about the average true diabetes cases in the spatial units.

$$\lambda \sim \text{Gamma}(k, \theta)$$

where α and β are the gamma distribution's shape and rate parameters, respectively. Suppose there is a belief that the true diabetes cases are expected to be small or have a lower mean value. In that case, the gamma prior to being chosen will have a smaller α parameter and a larger β parameter.

3. Results

The under-reporting of diabetes cases was assumed to follow a Poisson-Binomial distribution with parameters λ, which is the mean distribution of the true cases, and P, which is the probability that a diabetes case is reported in a given county. Unlike in the original Poisson-Binomial model where P was said to be constant across the different spatial units, the parameter P in the proposed improved model varied from one county to another depending on three covariates; illiteracy level, access to healthcare facilities, and poverty index.

The estimation of parameters for the model on under-reporting diabetes cases in Kenya was carried out using the Gibbs sampling Markov Chain Monte Carlo (MCMC) method. The Gibbs sampling MCMC is a powerful technique that allows us to estimate the model's parameters. In each iteration of the for-loop, the Gibbs sampler updates the values of the parameters that we had, which include $\lambda, \beta_0, \beta_1, \beta_2 and\ \beta_3$ based on the full conditional distributions. The number of iterations that were used in this case was 1000. The parameters were estimated as follows; $\lambda = 32998$, $\beta_1 = 0.504$, $\beta_2 = 2$ and $\beta_3 = 0.437$.

3.1 Diabetes reporting probabilities

The data was fitted on a logistic regression model to help determine the probability that a case is reported in a given spatial unit. Unlike in the original model, where the probability of reporting was constant, the probability of reporting for all 47 counties varied according to the three covariates, illiteracy level, access to healthcare, and poverty index in all 47 counties. The probabilities for the top and bottom 10 counties were as shown below:

Table 1: Table of disease reporting probabilities

Top 10 counties	Reporting probability	Bottom 10 counties	Reporting probability
Migori	0.9002423	Mombasa	0.7164098
Kisumu	0.8970270	Kilifi	0.7235266
Busia	0.8937203	Lamu	0.7305321
Vihiga	0.8903203	Garissa	0.7374244
Bomet	0.8868254	Mandera	0.7442022
Nakuru	0.8795435	Isiolo	0.7508642
Elgeyo Marakwet	0.8718608	Tharaka	0.7574092
Trans Nzoia	0.8678650	Kitui	0.7638363
Kiambu	0.8595564	Baringo	0.7701409
Kirinyaga	0.8552406	West Pokot	0.7763341

The probability that a case was reported in Migori County was higher than in the other 46 counties. The higher probability was attributed to the poverty index, illiteracy level, and access to healthcare in that county. The other counties with a higher probability of reporting include Kisumu, Busia, Vihiga, Bomet, Nakuru, Elgeyo-Marakwet, Trans Nzoia, Kiambu, and Kirinyaga counties with reporting probabilities ranging from 0.9 to 0.855. The higher probabilities indicate a greater likelihood of diabetes cases being reported in these regions.

On the other hand, the bottom 10 counties include Mombasa, Kilifi, Lamu, Garissa, Mandera, Isiolo, Tharaka, Kitui, Baringo, and West Pokot, with a reporting probability ranging from 0.716 to 0.776. The lower reporting probabilities suggest a relatively lower prevalence of diabetes cases or potential underreporting.

3.2 Efficiency of the model

Efficiency was conducted by comparing the original model, where the probability of reporting a case in a given spatial unit is held constant in all the units, and the improved model, where the probability of reporting varies across the spatial units and depends on the covariates illiteracy level, access to healthcare and poverty index. The deviance information criterion (DIC) method was used to compare the two models. The results were as shown below:

Table 2: Model efficiency comparison

Model	DIC
Original model	2276.9
Improved Model	2056.4

From the table, the original Poisson-Binomial model has a DIC value of 2276.9, while the improved model has a DIC value of 2056.4. The DIC values suggest that the improved model performs better than the original model. A lower DIC value for the improved model indicates that it better fits the data while being less complex than the original model.
Therefore, the improved model is preferred because it has a low DIC value. It is considered a better compromise between the model fit and complexity, making it a more suitable choice for estimating the underreporting of data than the original model.

3.3 Spatial distribution of underreported cases

The underreported cases from each count were mapped, and the resultant chart was as shown below. The results indicated that Nairobi County had the highest rate of underreporting followed by Mombasa County. It was also clear that a majority of the counties had less than 4000 underreported cases of diabetes. The cases in the top 10 and bottom 10 counties was as shown in table 3 below:

Table 3: Underreported cases

County	Least Cases	County	Highest cases
Lamu	1269	Nairobi	16816
Nyamira	1424	Mombasa	11734
Vihiga	1496	Kiambu	7726
Elgeyo Marakwet	1632	Kilifi	6405
Tana River	1657	Uasin Gishu	6265
Baringo	1740	Murang'a	6088
Taita Taveta	1789	Meru	5965
West Pokot	1908	Nyeri	5559
Samburu	1974	Nakuru	5551
Isiolo	1996	Kakamega	5431

The underreported cases were mapped in all 47 counties in Kenya and the map displaying the severity of underreporting was as shown in chart 1 below:

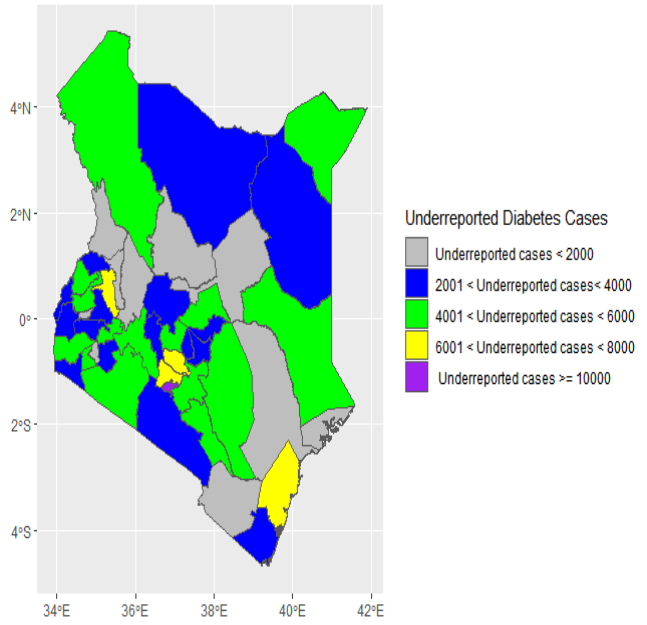

Figure 1: Distribution of underreported diabetes cases

3.4 Identification of high-risk Counties

Identifying counties with the highest reporting probabilities provides valuable insights into potential high-risk regions for diabetes in Kenya. Policy-makers and public health authorities can utilize this information to focus their efforts on targeted interventions and resource allocation. Despite having a high reporting rate, the top-performing counties in terms of reporting probabilities, such as Migori and Kisumu, still warrant increased attention for preventive healthcare measures, awareness campaigns, and the establishment of diabetes screening and treatment facilities. However, the counties with the least probability of reporting should be given priority than the top counties. This will be done to ensure that the ability of counties to report the cases of diabetes that occur is increased.

The underreported cases of diabetes from each of the 47 counties were added to the observed cases of diabetes for respective counties. This is important since it can be used to reveal the true cases of diabetes across the country. By incorporating underreported cases into our estimation, we can shed light on the healthcare disparities and inequities in different regions of Kenya. The true distribution of diabetes cases provides valuable insights into areas lacking healthcare resources and infrastructure. Understanding the variations in disease burden allows policymakers to direct their efforts and resources toward the most affected counties, thereby improving healthcare accessibility and outcomes. A table displaying the true diabetes count for the top and bottom counties is shown below:

Table 4: True cases the top and bottom counties

County	Highest cases	County	Least cases
Nairobi	179712	Lamu	7038
Kiambu	89789	Isiolo	11716
Nakuru	77039	Tana River	14295
Mombasa	60067	Samburu	14387
Meru	57794	Taita Taveta	15416
Machakos	54057	Marsabit	19435
Kisii	53857	Tharaka Nithi	20004
Kisumu	49190	Elgeyo Marakwet	20811
Kakamega	49134	Laikipia	21379
Bungoma	48017	West Pokot	22758

The true count of the diabetes cases for all the 47 counties in Kenya were used in making a chart. This chart is important since it will help the policymakers to know the true cases of diabetes in a given county and distribute enough resources that are meant to control the disease. The resulting map was as shown in chart 2 below:

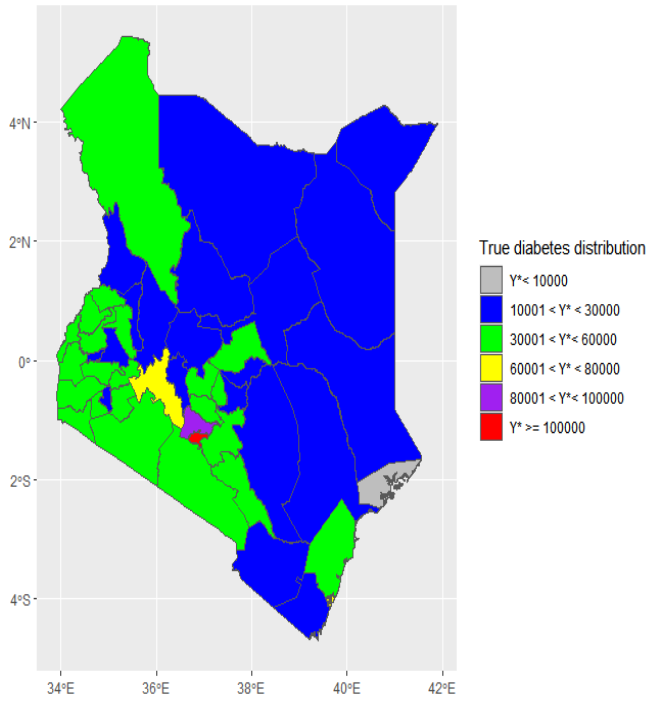

Figure 2: Distribution of diabetes cases in Kenya

From the distribution map, the county with the highest cases of diabetes in Kenya is Nairobi, with 179,604 cases. The second risky group of counties had cases of between 80,001 and 100,000, where Kiambu was the only county under that category. The third category comprises counties with cases between 60,001 and 80,000. This category had two counties, Nakuru with 76,968 cases and Mombasa with 60,067. The fourth group consisted of those counties with cases between 30,001 and 60,000. The counties in this category include Kisii with 53,894 cases, Bungoma with 48,056 cases, Uasin Gishu with 47,821 cases, Murang'a with 42,359 cases, Kwale with 39,492 and Kirinyaga with 37,786 counties, among others. The fifth category comprised counties with cases between 10,001 and 30,000. Some counties in this category include Kitui with 27,846 cases, Embu with 25,916 cases, Nyamira with 25,720 cases, Vihiga with 24,073 cases, and Tharaka with 20,004 cases, among others. The last category is those counties with less than 10,000 cases, with only 1 county, Lamu, with 7038 cases.

4. Conclusion

The spatial Bayesian analysis results provide valuable insights into the under-reporting of diabetes cases across the counties of Kenya. The research successfully implemented an improved Poisson-Binomial model, which considered the variation of reporting probabilities across spatial units based on covariates such as illiteracy level, access to healthcare, and poverty index. The results demonstrated that the improved model outperformed the original model, offering a better fit to the data while being less complex. This finding emphasizes the importance of accounting for spatial heterogeneity in disease reporting when estimating the true distribution of diabetes cases.

To enhance the accuracy and reliability of diabetes reporting, future efforts should prioritize addressing underreporting through investments in robust data collection mechanisms, community health outreach programs, and training healthcare professionals in accurate case reporting. By doing so, a comprehensive understanding of the true distribution of diabetes cases in Kenya can be achieved, allowing for more effective and targeted public health strategies to combat diabetes and improve the population's overall health.

References

[1] Ayugi, B., Nyongesa, M. K., & Ondieki, M. (2019). Spatial Analysis of Prevalence and Factors Associated with Underreporting Diabetes Mellitus in Kenya. International Journal of Environmental Research and Public Health, 16(22), 4527.

[2] Kenya Diabetes Study Group, (2019). Improving Diabetes Care at Primary Healthcare Level.

[3] Koch, T. (2005). Cartographies of Disease: maps, mapping, and medicine. Esri Press Redlands, CA.

[4] Manda, S. O., & Feltbower, R. G. (2018). Spatial modeling of under-reporting of notifiable infectious disease counts. Journal of the Royal Statistical Society: Series A (Statistics in Society), 181(4), 1099-1121.

[5] Moore, D. A., Carpenter, T. E., et al. (1999). Spatial analytical methods and geographic information systems: use in health research and epidemiology. Epidemiologic reviews, 21(2).

[6] Mugendi, B., Amugune, B., & Aluoch, J. R. (2019). Analysis of Spatial Variation and Under-Reporting of Cholera Cases in Kenya. Applied Spatial Analysis and Policy, 12(3), 561-576.

[7] Moraga, P. and Lawson, A. B. (2012). Gaussian component mixtures and car models in Bayesian disease mapping. Computational Statistics & Data Analysis, 56(6).

[8] Neubauer, G., Djuraš, G., and Friedl, H. (2016). Models for underreporting: A Bernoulli sampling approach for reported counts. Austrian Journal of Statistics, 40(1&2).

[9] Ngesa, O., Achia, T., and Mwambi, H. (2014a). A flexible random effects distribution in disease mapping models. South African Statistical Journal, 48(1).

[10] Ngesa, O., Mwambi, H., and Achia, T. (2014b). Bayesian spatial semi-parametric modeling of HIV variation in Kenya. PloS one, 9(7).

[11] Oti-Boateng, E., Ngesa, O. and Osei, F. (2016). Bayesian disease mapping in presence of under-reporting.

[12] WHO (2014). Risk predictive modeling for diabetes and cardiovascular disease. Bulletin of the World Health Organization, 51(1).

[13] Winkelmann, R. (1996). Markov chain Monte Carlo analysis of underreported count data with an

application to worker absenteeism. Empirical Economics, 21(4).

[14] Winkelmann, R. and Zimmermann, K. F. (1993). Poisson-logistic regression. Volkswirtschaftl. Fakultät d. Ludwig-Maximilians-Univ. München.

[15] Yang, S., Zhao, Y., and Dhar, R. (2010). Modeling the underreporting bias in panel survey data. Marketing Science, 29(3).

[16] Ye, F. and Lord, D. (2011). Investigation of effects of underreporting crash data on three commonly used traffic crash severity models: multinomial logit, ordered probit, and mixed logit. Transportation Research Record: Journal of the Transportation Research Board, (2241).

[17] Zayeri, F., Salehi, M., and Pirhosseini, H. (2011). Geographical mapping and Bayesian spatial modeling of malaria incidence in Sistan and Baluchistan province, Iran. Asian Pacific journal of tropical medicine, 4(12).