BAYESIAN MUTINOMIAL ORDINAL MODEL TO ANALYSE THE SPATIAL PATTERNS OF CHILDHOOD ANAEMIA IN TANZANIA

Rasheed Adeyemi, Temesgen Zewotir & Shaun Ramroop
School of Mathematics, Statistics and Computer Science,
University of KwaZulu-Natal, South Africa
e-mail: *Corresponding author: adeyemira@yahoo.ca*

Key words: Tanzania, aneamia, Under-five children, Semiparametric model, Spatial effect

Summary: This paper used self-reported data obtained from 2010 Tanzania Demographic and Health Survey. We propose a semi-parametric model that combines the fixed effects, non-linear term and spatial component in a unified framework. The fixed(linear) effects were modelled parametrically, and the non-linear effects of metrical covariates were modelled using P-splines. The spatial effect was modelled using a Markov random field prior. We explore multinomial logit models to analyse the severity of anaemia among under-five children and assess the risk factor of childhood anaemia. We run several Bayesian models via Markov Chain Monte Carlos(MCMC)simulation techniques and the models were compared using Deviance information criteria(DIC). We found the spatial residual pattern of anaemia and the risk factors. The risk factors associated with anaemia include place of residence, maternal poverty index, childhood under-nutrition, and vitamin A supplementation and infectious diseases. The findings also estimate non-linear function of continuous covariates(child's age and maternal body mass index (mbmi)) on childhood anaemia. Our method also estimate the residual spatial effects that are not captured by the underlying factors and produced probability predictive maps. Higher residual risk were identified in Northern-South of Tanzania. These spatial maps highlight high endemic regions, that can assist government agency to target scarce health resource and effective policy making.

1. Introduction

Childhood anemia is a global public health problem, with monumental consequences at adulthood. It relates to major causes of health problems in children, and adversely affects their cognitive and physical development(Denny, Kuchibhatla and Cohen, 2006), and immunity, increases the risk of infections and infant mortality(Organization et al., 2008). A WHO recent report on the world prevalence of anemia, showed that the global prevalence of anaemia is 24.8% with the highest prevalence in sub-Saharan Africa (67%), followed by the southern east Asian(65.5%).

Several studies have identified genetic determinants (Meinzen-Derr et al., 2006; **?**), socio-economic, cultural and dietary(Hadler, Colugnati and Sigulem, 2004; **?**) factors on anaemia using linear and binary logistic regression model. However, little work has been done on investigating the geographical variations and other underlying factors that are not linearly associated with child's anaemia. The motivation of this work is to provide a flexible approach that simultaneously estimate linear and non-linear covariateas , as well as small area estimation of spatial heterogeneity across the regions (districts).

2. Material and Methods

2.1. Model formulations

In 2010 Tanzania Demography Health Survey (TDHS)data, child‘s anemia defined as a measure of haemoglobin concentration and extracted with aim of assessing the influence of some covariates on the childhood anaemia. TDHS data set contains several other variables, only those that are related to anemia level and those similar to the ones identified in the literate were selected. The children involved in the survey had ages range between 1- 59 months and the respondents(mothers) are in reproductive ages range 15- 49.
Model A: Anaemia ia a product of low level of functional Hb in the blood. Hence, the concentration of Hb in the blood was considered as continuous variable, i.e. $y_{ij}$and modelled by assuming a Gaussian distribution.
Model B: The response variable, anaemia can be classified by the haemoglobin concentration level(Hb) as

$$y_{i1} = \begin{cases} 1 & \text{if Hb concentration level of a c hild is} \leq 11.0 \text{ g/dl} \\ 0 & \text{otherwise} \end{cases}$$

Model C : The severity level of anemia in child can vary based on the concntration of Hb level. According to Who Health organization (Organization et al., 1968), Hb can be classified as severe, moderate, mild or normal resulting in a four- ordered category and the response variable, y_{ij} can be constructed as

$$y_{i2} = \begin{cases} 1: & \text{non-anemia, if Hb} \geq 11.0\text{g/dL} \\ 2: & \text{mild anemia, if } 10.0 \text{ g/dL} \geq Hb \leq 10.9 \text{ g/dL} \\ 3: & \text{moderate anemia, if } 7.0 \text{ g/dL} \leq Hb \leq 9.9 \text{ g/dL} \\ 4: & \text{severe anemia, if } Hb < 7.0 \text{ g/dL} \end{cases}$$

where y_{i1} is a univariate response (continuous, binary response outcome) and y_{i2} is an ordered categorical response outcome.
The present study intends to apply a flexible regression model to quantify the fixed and non-linear effects, as well as geographical variations of the anaemia level as response variables y_{i1} and y_{i2} as defined above.

2.2. Multinomial ordinal models

The ordinal logistic regression model can be expressed as a latent variable model (Agresti, 2003; Tutz, 2011). These models are regarded as a powerful class of models for for treating observations that fall into mutually categorical classes. The ordered nature facilitates the use of a flexible regression framework, which allows better inference. A regression model based on multi-categorical outcomes sometimes called cumulative regression models had been earlier investigated in literature

(McCullagh and Nelder, 1989; **?**). The later authors argued that the cumulative link models are members of a class of multivariate generalized linear models.
The models can be motivated from latent variables such that the response variable Y, here, anaemia concentration, is a categorized version of a continuous latent (utility) variable defined by

$$Z = \eta + \varepsilon \tag{1}$$

where η is a predictor depending on covariates and parameters and ε is the error term. The two variables Y and Z are linked by $Y = j$ if and only if

$$\theta_{j-1} < Z \leq \theta_j, \quad j = 1,2,3,4 \tag{2}$$

with thresholds $-\infty < \theta_0 < \theta_1 < \ldots < \theta_k = \infty$. In a multinomial logit model setting, the error variables ε in (1) are independent across the categories and assumed to be standard extreme value distributed with function F. Hence, it follows that Y obeys a cumulative logit model. The predictor is then defined as

$$Pr(y_i \leq j|\eta) = F(\theta_j - \eta) \tag{3}$$

If F in equation (3)is chosen to be the logistic distribution function, one obtains a sequential logit model (Tutz, 2003). Hence, the j^{th} the level of anaemia is estimated as the probability of selecting that category against the reference, in our case, severe anaemia is chosen. The influence of covariates is modelled using the multinomial logit model given as

$$Pr(y_i = j|y_i \geq j, \eta_i) = \frac{\exp(\eta_j)}{1 + \exp(\eta_j)} = \theta_j - \eta_j \tag{4}$$

If F is chosen to be the logistic distribution function, one obtains the proportional odds model

$$\log \frac{Pr(y_i \leq j|\eta_i)}{Pr(y_i > j|v_i)} = \theta_j - \eta_i \tag{5}$$

When the logit link is replaced by the complimentary log-log link, the resulting model is written as

$$\log \left[\log \frac{Pr(y_i \leq j|v_i)}{Pr(y_i > j|v_i)} \right] = \theta_j - v_i'\gamma \tag{6}$$

Equation (5) is known as the proportional hazard model (Tutz, 1991; **?**)

2.3. Bayesian structured additive regression model

Consider a set of regression observations $(y_i, x_i, s_i, v_i), i = 1, 2, \ldots, 4870$, where y_i is either binary or categorical response variable, a vector x_i is the of metrical covariate effects of the mother‘s age at birth and body mass index, the spatial covariate $s_i \in [1, \ldots, S]$, index of the district(region) where mother i lives in Tanzania and a further vector $\mathbf{v} = (v_{i1}, \ldots, v_{iq})$ of categorical covariates.
For a linear case,the linear predictor is written as

$$\eta = \gamma v_1 + \ldots + \gamma_q v_q = \theta_j - \mathbf{v}'\gamma \tag{7}$$

where $\gamma = (\gamma_1, \ldots, \gamma_q)'$ is unknown and must be estimated together with the unknown thresholds,θ_j from the data. For identifiability, the linear combination does not contain an intercept term γ_0, otherwise one of the threshold must be set to zero. Usually, the last category is chosen to be equal to zero.

Equation (7) can be modified to include geo-reference of the woman, where the mother i lives. Thus, the regression model prediction η now called a geoadditive predictor for incorporating the geographical location for a particular woman i, and the semiparametric predictor by (Tutz, 2003) is given as

$$\eta_i = \theta_i^j - f(x_i) + f_{spat}(s_i) + v_i'\gamma \tag{8}$$

where, $f(x_i)$,$f_{spat}(s_i)$ and v_i represent the estimates of the unknown non-linear smoothing effects of the metrical covariates $x_i$such as mother‘s age at birth, the spatial effect and a vector of the fixed effect parameters. The spatial component, $f_{spat}(s_i)$ of the model can be used to capture the random effects of area $s_i, , s \in \{1, \ldots, 37\}$, where the woman$i$ resides. The spatial component, $f_{spat}(s_i)$ is further split into two components:$f_{str}(s_i)$ and $f_{unstr}(s_i)$ as spatially structured (correlated) and unstructured(uncorrelated) random effects respectively.

A univariate response variable with a non-linear predictor is defined by

$$\eta_i = v_i'\gamma + f_{mbmi}(mbmi) + f_{cage}(cage) + f_{district}(s_i) \tag{9}$$

For the ordered categorical response variable y_{i2}, cumulative probit models with predictor similar to (9) were fitted as

$$\eta_i^j = \theta_i^j - (v_i'\gamma + f_{mbmi}(mbmi) + f_{cage}(cage) + f_{district}(s_i)) \tag{10}$$

In equation (10), the smooth effect functions of the nonlinear effects aref_{cage} (child‘s age, in months), f_{mbmi} (mother‘s body mass index $= weight(kg)/height^2(meters)$, and $f_{spat}(s_i)$ are the structured spatial effects, index $s_i \in \{1, \ldots, S\}$. The spatial effects can be split into structured (correlated) and unstructured effects are used to capture any residual spatial variation between and within districts (regions) that are not explained by the other underlying determinants of anaemia in the model. In this application, $y_{i2} = 4$ is chosen as the reference category. For the multinomial model, the covariates are assumed to be independent of the category while the effects are category-specify. For the ordinal model, all effects apart from the thresholds are independent across categories.

2.4. Bayesian inference

Prior distributions

Within a Bayesian framework, all model parameters and nonlinear functions are usually taken as random variables and an appropriate prior is needed to be specified for each. we need to specify appropriate prior for the model parameters. For the fixed regression parameters, γ‘s, a suitable choice is the independent diffuse prior, i.e. $p(\gamma) \propto constant$.

For the non-linear effect, a Bayesian P$-$splines prior was assumed as suggested by (Brezger, 2005). The P$-$spline permits for non-parametric estimation of f as a linear combination of the basis function (**B**$-$spline):

$$f(x) = \sum_{j=1}^{p} \xi_{kj} \mathbf{B}_j(x)$$

where$\mathbf{B}_j(x_{ij})$ are $\mathbf{B}-$ spline basis functions and $\xi = (\xi_1, \ldots .\xi_p)\prime$ correspond to the vector of the unknown regression coefficients. To achieve the smoothness of function$f$, we penalize the differences of coefficients of the adjacent $\mathbf{B}-$ splines as proposed by (Marx and Eilers, 1998). They suggest a moderate number of knots, such like between 20 to 40 nots and by introducing a roughness penalty on adjacent regression coefficients that regularize the smoothness to avoid overffiting. The coefficients was later replaced with Bayesian smoothing splines Hastie and Tisbshirani, 2000) or a flexible first and second order random walk as proposed by (Fahrmeir and Lang, 2001), defined by

$$\xi_j = \xi j - 1 + u_j; \qquad \xi_j = 2\xi_{j-1} - \xi_{j-2} + u_j$$

with Gaussian errors $u_j \sim N(0; \tau^2)$ and non-informative prior, $\beta_1, \beta_2, \ldots \propto const$. Again, τ_j^2 controls the smoothness of f. The variance parameter with hyperparameters a and b has inverse gamma distribution i.e ($\tau^2 \sim IG(a,b)$), and by assigning large (small) variance leading to less smoothing (smoother) on the curve.

In order to capture spatial effects, one chooses a Gaussian Markov random field prior which is common in spatial statistics, see (Besag, York and Mollié, 1991) and the unstructured spatial random effects, $f_{unstr}(s)$, takes exchangeable normal priors, $f_{unstr} \sim N(0, \tau^2_{unstr})$, where τ^2_{unstr} is a variance component that allows for over-dispersion and spatial heterogeneity.

For regions that exhibit geographical variation, we modelled with a structured spatial effect, f_{str}, which assumes that two sites or regions s and t are neighbours if they share a common boundary information. The structured spatial effects is then specified by the conditional autoregressive CAR error

$$f_{str}(s)|f_{str}(t), t \neq s, \tau^2 \sim N\left(\sum_{t \in \theta_s} \frac{f_{str}(t)}{N_s}, \frac{\tau^2}{N_s}\right) \tag{11}$$

where N_s is the number of adjacent regions and $t \in \theta_s$ denotes that the region t is a neighbour of region s. Thus, the conditional mean of $f_{str}(s)$ is an unweighed average of function evaluations for neighbouring regions. Spatial correlation between regions is achieved by introducing suitable spatial correlation structure on f_{str} in (11). This is specified by using either the Markov Random field(MRF) or Gaussian RF prior. The MRF is defined by

$$f_{str}(s)|\tau^2 \sim N\left(0, \tau^2_{str} Q^{-1}\right) \tag{12}$$

The density of vector f_{str} has mean zero and precision matrix Q defined as

$$Q_{st} = \begin{cases} m_s & s = t \\ -1 & s \sim t \\ 0 & \text{otherwise} \end{cases} \tag{13}$$

where $s \sim t$ denotes that region s is adjacent to t, m_s is the number of adjacent regions to s. Other options for modeling spatial effects such like (11) models and stationary Gaussian random field (Kriging) models can be found in (Rue and Held, 2005).

3. Data application and Results

The output of the analysis are presented in terms of tables of fixed effects of categorical covariates, the nonlinear plots of continuous covariates and the residual plots of spatial effects. Because of page limitation in this proceeding, we have only presented the results for Gaussian and multinomial models in this section.

Fixed effects

Table 1: Posterior mean and 95% credible interval of fixed Effects of categorical covariates

	Gaussian model			Multinomial logit model		
Variable	Mean	STD	95% cred.int	Mean	STD	95% cred.int
Intercept	6.248	4.274	(0.638 , 13.845)	-	-	-
θ_1	-	-	-	5.874	-	(- , -)
θ_2	-	-	-	6.809	-	(-, -)
θ_3				8.948	-	(- , -)
ref. No	0	0	0	0	0	0
stunted(HAZ)	0.013	0.070	(-0.116 , 0.156)	-0.096	0.101	(-0.293, 0.101)
wasted (WHZ)	-0.058	0.072	(-0.202, 0.081)	0.098	0.097	(-0.091, 0.287)
underweight(WAZ)	-0.045	0.076	(-0.198 , 0.097)	0.142	0.110	(-0.073, 0.357)
ref. Antenatal visit 5 +	0	0	0	0	0	0
1-3	-0.105	0.089	(-0.289 , 0.069)	0.316	0.138	(0.045, 0.586)
4-5	0.058	0.092	(-0.121, 0.250)	0.008	0.141	(-0.268, 0.285)
ref. No	0	0	0	0	0	0
Iron syrup dur. preg.	0.109	0.146	-0.173, 0.393	-0.078	0.218	(-0.506, 0.350)
ref. No	0	0	0	0	0	0
Vitamin A	0.050	0.060	(-0.074, 0.167)	-0.106	0.083	(-0.269, 0.057)
ref. Rural	0	0	0	0	0	0
urban	-0.062	0.086	(-0.224, 0.100)	0.024	0.122	(-0.216, 0.264)
ref. Female	0	0	0	0	0	0
Male	0.017	0.056	(-0.089, 0.124)	-0.001	0.079	(-0.156, 0.154)
ref. incomplete prim	0	0	0	0	0	0
prim	0.203	0.191	(-0.165, 0.561)	-0.188	0.308	(-0.791, 0.416)
sec	-0.334	0.201	(-0.747, 0.030)	0.514	0.324	(-0.121, 1.149)
high	0.221	0.512	(-0.877, 1.136)	-0.557	0.866	(-2.253, 1.140)
ref. poorest	0	0	0	0	0	0
poor	0.007	0.112	(-0.218, 0.218)	0.012	0.175	(-0.330, 0.354)
middle	-0.018	0.122	(-0.283, 0.211)	0.232	0.161	(-0.084, 0.549)
richer	-0.147	0.103	(-0.345 ,0.0625)	0.105	0.153	(-0.195, 0.405)
richest	-0.334	0.166	(-0.668, -0.014)	0.353	0.245	(-0.127, 0.833)
ref. $\leq 1child$	0	0	0	0	0	0
≥ 2	-0.041	0.098	(-0.234 , 0.139)	0.135	0.132	(-0.124, 0.394)
ref. - Flush toilet	0	0	0	0	0	0
latrine	0.220	0.092	(0.038, 0.402)	-0.301	0.125	(-0.545, -0.056)
Bush/field	-0.214	0.125	(-0.444 , 0.025)	0.228	0.172	(-0.109, 0.566)
ref. No disease	0	0	0	0	0	0
diarrhea	0.100	0.066	(-0.026 , 0.223)	-0.135	0.094	(-0.320, 0.050)
cough	6.054	4.26	(-1.481, 11.528)	4.571	-	(-, -)
fever	-0.173	0.059	(-0.287, -0.047)	0.209	0.089	(0.034, 0.383)
pneumonia	0.016	0.055	(-0.086, 0.120)	-0.022	0.084	(-0.186, 0.142)

ref. - reference category; − = not estimable

Non-linear effects

The estimates of non-linear function of child‘s age(in months) and mother‘s body mass index for Gaussian and multinomial model are presented in Figure 1. Each non-linear graph consists of a center line representing the posterior mean estimate bounded by 95% credible intervals(inner lines) and 80% credible interval(outer lines).

Spatial effects

Figure 1 showing the posterior means *(left panel)* and 95% credible interval *(right panel)*, which is used to determining the significance level of spatial variations. Black coloured regions are associated with low prevalence of childhood anaemia , white “coloured” regions depict high prevalence of anaemia, and grey coloured regions indicate, although these variations are not significant in this report.

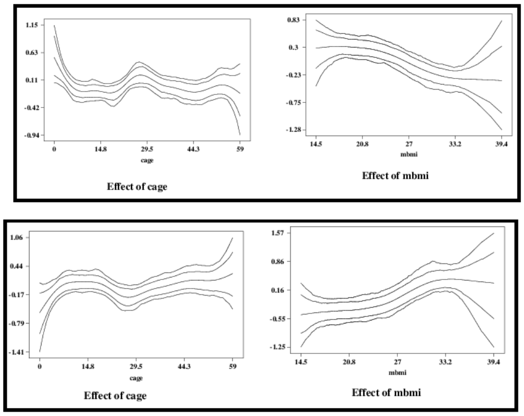

Figure 1: Nonlinear effects of child age and mother body mass index for Gaussian model *upper panel* and Multinomial logit model *lower panel*

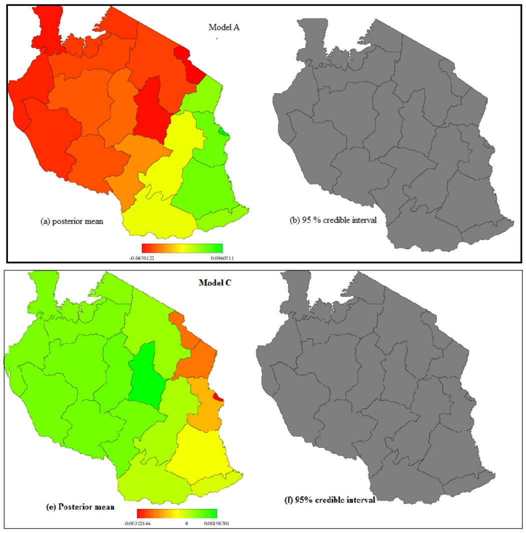

Figure 2: Spatial structured residual effects for Gaussian model *upper panel* and Multinomial logit model *lower panel*

4. Conclusion

The paper investigates the impacts of different kinds of covariates on the childhood anaemia. Our approach is flexible and robust, and estimate several effects simultaneously. In addition to the statistical relevance of the output, we produce spatial residual effects which may be neglected in classical regression settings. The spatial residual maps can assist developing partners and government agents to channel health resources in a more effective manner.

Acknowledgement

We would like to acknowledge permission granted by MEASURE DHS to use the 2010 Tanzania DHS data under the project- Spatial Analysis of poverty, malnutrition and childhood mortality in Sub-Saharan Africa. The first author also appreciate the study fellowship received from Federal University of Technology, Minna-NIGERIA.

References

AGRESTI, A. (2003). Logit models for multinomial responses. *Categorical Data Analysis, Second Edition*, 267–313.

BESAG, J., YORK, J., AND MOLLIÉ, A. (1991). Bayesian image restoration, with two applications in spatial statistics. *Annals of the institute of statistical mathematics*, **43** (1), 1–20.

BREZGER, A. (2005). *Bayesian P-splines in structured additive regression models*. Ph.D. thesis, lmu.

DENNY, S. D., KUCHIBHATLA, M. N., AND COHEN, H. J. (2006). Impact of anemia on mortality, cognition, and function in community-dwelling elderly. *The American journal of medicine*, **119** (4), 327–334.

FAHRMEIR, L. AND LANG, S. (2001). Bayesian semiparametric regression analysis of multicategorical time-space data. *Annals of the institute of Statistical Mathematics*, **53** (1), 11–30.

HADLER, M.-C. C., COLUGNATI, F. A., AND SIGULEM, D. M. (2004). Risks of anemia in infants according to dietary iron density and weight gain rate. *Preventive medicine*, **39** (4), 713–721.

MARX, B. D. AND EILERS, P. H. (1998). Direct generalized additive modeling with penalized likelihood. *Computational Statistics & Data Analysis*, **28** (2), 193–209.

MCCULLAGH, P. AND NELDER, J. A. (1989). *Generalized linear models*, volume 37. CRC press.

MEINZEN-DERR ET AL. (2006). Risk of infant anemia is associated with exclusive breast-feeding and maternal anemia in a mexican cohort. *The Journal of nutrition*, **136** (2), 452–458.

ORGANIZATION, W. H. ET AL. (1968). Nutritional anaemias: report of a who scientific group [meeting held in geneva from 13 to 17 march 1967].

ORGANIZATION, W. H. ET AL. (2008). Worldwide prevalence of anaemia 1993-2005: Who global database on anaemia.

RUE, H. AND HELD, L. (2005). *Gaussian Markov random fields: theory and applications*. CRC Press.

TUTZ, G. (1991). Sequential models in categorical regression. *Computational Statistics & Data Analysis*, **11** (3), 275–295.

TUTZ, G. (2003). Generalized semiparametrically structured ordinal models. *Biometrics*, **59** (2), 263–273.

TUTZ, G. (2011). *Regression for categorical data*, volume 34. Cambridge University Press.