Quantifying Disease Burden in Low-Resource Settings: Statistical Insights and Approaches with Model-Based Geostatistical Modelling.

Chipeta M G[*,1], Masoambeta J[2] and Khundi M[1]

[1] African Institute for Development Policy (AFIDEP)
[2] National Statistical Office (NSO)
[*]Correspondence: Michael.chipeta@afidep.org

Introduction: It is crucial to quantify the disease burden in low- and middle-income countries (LMICs) to inform public health efforts and optimise resource allocation. However, LMICs frequently encounter obstacles in accessing comprehensive and dependable disease registries, which hinders accurate disease burden estimation. Even when the registers are present, they may present biased estimates because of barriers to access to care. Populations with better access might be over-represented compared to those that do not have good access to care. This paper reviews and applies statistical insights and approaches to disease modelling using Generalised Linear Geostatistical Modelling (GLGM), an extension of Generalised Linear Models (GLM), overcoming data limitations and providing valuable insights into the disease burden in LMICs. The methods are applied to the mapping of anaemia prevalence in children under five in Malawi.

Methods: GLGMs provide a robust framework integrating statistical modelling techniques with geospatial analysis to generate precise disease burden estimates. Without comprehensive disease registries, GLGMs can offer valuable insights into disease patterns and risk factors by integrating available data sources and incorporating spatial correlation. This methodology is essential for LMICs, where disease registries are frequently incomplete and of limited availability and quality.

GLGMs permit the incorporation of alternative data sources, such as hospital records, community/population surveys, and demographic data, to estimate disease burden without complete disease registries. GLGMs enable a comprehensive comprehension of disease distribution and burden at regional and local levels (i.e., high spatial resolutions of up to 1 by 1 km, data permitting) in LMICs by utilising spatial information and accounting for covariates such as demographic factors and environmental variables.

In the current analysis, we fit a GLGM to survey data to model the prevalence of anaemia in children under five using demographic and health survey (DHS) data from the 2015-16 Micronutrient Survey, a sub-study of the Malawi Demographic Health Survey

(MDHS). We let the design $X = x_1, \ldots, x_n$ denote a set of n distinct spatial locations representing the geographical coordinates (i.e., longitudes and latitudes) of sampled households, villages or communities in the prevalence survey. At each location x_i, we then sample m_i individuals and perform a test for the disease outcome of interest. Conditionally, on a spatial stochastic process $S(x_i)$and mutually independent zero-mean Gaussian latent variables Z_i, we assume that Y_i are mutually independent binomial variables with a probability of having a positive outcome p_i. A logit link function is then used for p_i, assuming the form:

$$log\left[\frac{p(x)}{1-p(x)}\right] = d(x_i)'\beta + S(x_i) + Z_i \qquad \text{Equation 1,}$$

where the vector $d(x_i)$ contains explanatory variables that are frequently acquired from remotely sensed images (e.g., population density, rainfall, temperature, and NDVI) or that pertain to individual households (e.g., age, education and socioeconomic status). The elements of the vector β represent regression coefficients associated with each of these covariates. To account for unmeasured spatially structured risk factors that generate residual spatial correlation among observations, the spatial random effect $S(x_i)$ is applied. The unstructured residuals Zi, which are frequently called the "nugget effect", can be interpreted in two different ways: as extra-binomial variation within households (e.g., genetic variation) or as small-range spatial variation (on a range shorter than the observed minimal distance between locations). From the model in Equation 1, we generate summary statistics, as well as measures of uncertainty and other metrics, such as exceedance probabilities, to identify hotspots.

Results and discussion: The implementation of GLGM in LMICs has numerous benefits. It allows for identifying disease hotspots, evaluating intervention strategies, and prioritising scarce resources based on accurate burden estimates (i.e., towards areas/populations that need the most resources, thereby playing a critical role in advocating equitable access to resources). In addition, GLGM permits the investigation of complex relationships between disease outcomes and various covariates, thereby shedding light on the underlying drivers of disease burden in LMICs.

Children under 5 underwent anaemia testing, with 4,601 included in the study. The prevalence of child anaemia in Malawi is reported at 62.7%. Key factors include the child's age, with older children exhibiting a lower likelihood of anaemia. Those with a fever are more likely to have anaemia. Higher levels of maternal education (secondary and above) and greater household wealth (middle and above) were found to be protective against anaemia. Additionally, children from urban settings showed a lower likelihood of having anaemia.

In **Figure 1**, we depict the geospatial framework for the GLGM process from model inputs such as input data, geospatial covariates and the model implementation/process, resulting in model summary statistics and measures of uncertainty.

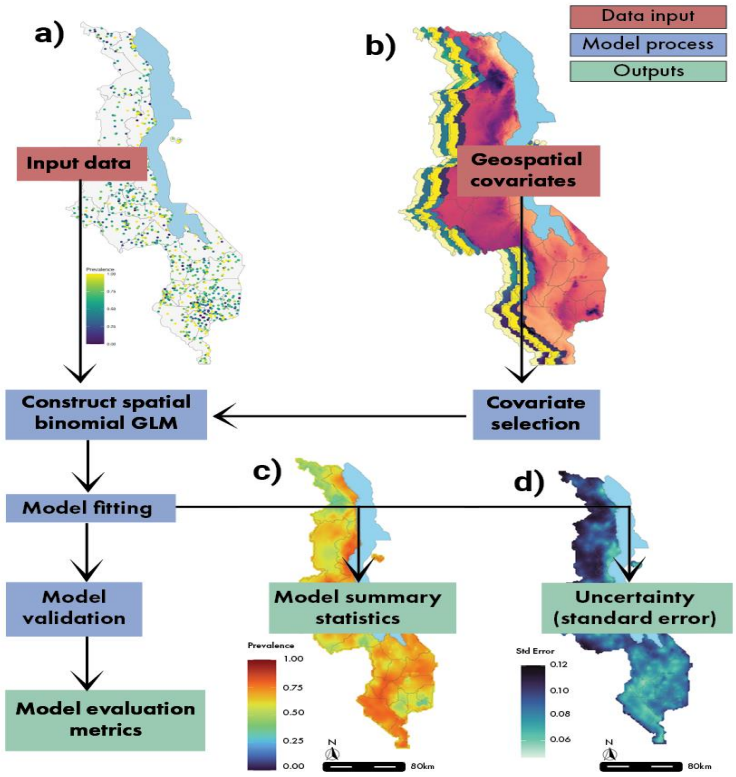

Figure. 1: *Flowchart detailing model construction, fitting, and validation process. (a) Input from survey data indicates child anaemia point prevalence at the cluster/enumeration area level. (b) Geospatial covariates at 5 x 5 km resolution. (c) Prediction (mean) surface for child anaemia at 5 x 5 km resolution. (d) Uncertainty (standard error) for child anaemia at 5 x 5 km resolution.*

Geographically, the prevalence of child anaemia is homogeneously high throughout Malawi, with pockets of very low prevalence and very high (i.e., hotspots), as exhibited by the 95% confidence interval maps (to be shown in the presentation), with the highest quantile placing the country at over 80% prevalence. The northern and southern highlands of Malawi exhibit the lowest prevalence rates. The hotspots are identified in the low-lying lakeshore districts (Karonga, Nkhata-Bay, Nkhota-kota, Salima, and Mangochi), as well as in the southern districts of Chikwawa, Nsanje, and certain areas of Mwanza and Neno; see **Figure 2**. These hotspot areas are places where prevalence exceeds 65%, exceeding the national prevalence average. Urban settings also exhibit low anaemia prevalence. The model's uncertainty measure will be presented using standard errors, depicting low standard errors close to data collection clusters.

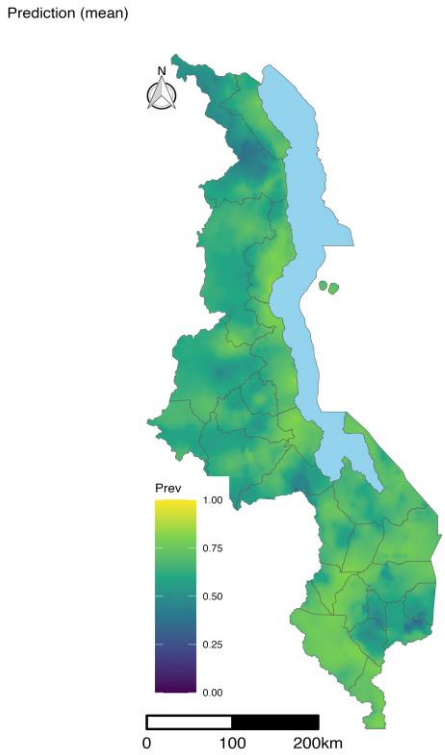

Fig. 2 Anaemia prevalence predictions among children aged under five years in Malawi.

Conclusion: This paper highlights the significance of statistical insights and approaches, particularly GLGM, in quantifying disease burden in LMICs, where comprehensive disease registries are frequently unavailable or incomplete. By leveraging available data sources (such as surveys, i.e., DHS) and integrating spatial information, GLGM provides policymakers and public health practitioners with a valuable tool for understanding disease burden, allocating resources efficiently, and designing targeted interventions in LMIC settings. We have applied the GLGM methodology to map the prevalence of anaemia in children under five in Malawi using 2015/16 demographic and health survey data.