Kandala, Fahrmeir, Klasen:

Geo-additive models of Childhood Undernutrition in three Sub-Saharan African Countries

Sonderforschungsbereich 386, Paper 287 (2002)

Geo-additive models of Childhood Undernutrition in three Sub-Saharan African Countries

Ngianga-Bakwin Kandala*, Ludwig Fahrmeir*, Stephan Klasen**
University of Munich, Germany.

Abstract

We investigate the geographical and socioeconomic determinants of childhood undernutrition in Malawi, Tanzania and Zambia, three neighboring countries in Southern Africa using the 1992 Demographic and Health Surveys. We estimate models of undernutrition jointly for the three countries to explore regional patterns of undernutrition that transcend boundaries, while allowing for country-specific interactions.

We use semiparametric models to flexibly model the effects of selected socioeconomic covariates and spatial effects. Our spatial analysis is based on a flexible geo-additive model using the district as the geographic unit of analysis, which allows to separate smooth structured spatial effects from random effect. Inference is fully Bayesian and uses recent Markov chain Monte Carlo techniques.

While the socioeconomic determinants generally confirm what is known in the literature, we find distinct residual spatial patterns that are not explained by the socioeconomic determinants. In particular, there appears to be a belt running from Southern Tanzania to Northeastern Zambia which exhibits much worse undernutrition, even after controlling for socioeconomic effects. These effects do transcend borders between the countries, but to a varying degree.

These findings have important implications for targeting policy as well as the search for left-out variables that might account for these residual spatial patterns.

Keywords : Sub-Saharan African Countries; Geo-additive models; undernutrition; spatial statistics; semiparametric Bayesian analysis.

1 Introduction

Childhood undernutrition is among the most serious health issues facing developing countries. It is an intrinsic indicator of well-being, but is also associated with morbidity, mortality, impaired childhood development, and reduced labor productivity (Sen, 1999; UNICEF, 1998; Pritchett and Summers, 1994; Pelletier, 1998, Svedberg 1999). Reducing malnutrition rates by half is one of the central development goals adopted by the international community at the Millennium Summit (UN, 2000).

There is a sizeable theoretical and empirical literature on the determinants of childhood undernutrition in developing countries (see Smith and Haddad, 1999, 2001, and UNICEF, 1998 for a survey). Most studies use parametric approaches to modelling the socioeconomic determinants of undernutrition.

In this paper, we present two innovations on this literature. First, we use flexible regression methods to model the effects of covariates that clearly have nonlinear effects on stunting. Secondly, we use flexible methods to modelling spatial determinants of undernutrition and allocate these spatial effects to structured and unstructured (random) components. This is done jointly in one estimation procedure that thereby simultaneously identifies socioeconomic determinants, and the spatial effects that are not explained by these socioeconomic determinants. In this way, we are able to identify regional patterns of undernutrition that are either related to left-out socioeconomic variables that have a clear spatial pattern or point to spatial (possibly epidemiological) processes that account for these spatial patterns. Identifying spatial patterns of undernutrition beyond the known socioeconomic determinants should also assist in poverty mapping and associated regional targeting of resources (Elbers *et.al.*, 2000).

We apply these methods to an analysis of chronic undernutrition (stunting) in Malawi, Tanzania, and Zambia using the 1992 Demographic and Health Surveys (DHS). Malawi, Tanzania and Zambia are neighboring low- income countries in Southern Africa, all belonging to the poorest countries in the world, with very poor education, health, and human development indicators. They have been affected by years of economic stagnation and decline as shown by negative per capita growth rates throughout the 1980s and early 1990s, and have also experienced deteriorations in health and education indicators (World Bank, 2000, see Table 1.1). More recently, they have been severely affected by the HIV-AIDS pandemic. Stunting is a serious problem in all three

countries, affecting some 48.7% of children in Malawi, 46.7% in Tanzania, and 39.6% in Zambia.

Table 1.1 Socioeconomic Data for Malawi, Tanzania, and Zambia for 1992 .[1]

	Malawi	Tanzania	Zambia
GNI p.c. 1992	470	450	750
Growth 80-92	-2.0	-2.5	-3.2
Life Expectancy	44	49	49
Sec Enrol (% gross)	9.2	5.3	26
HDI 1992	0.33	0.364	0.425
HIV Prevalence 94	13.6	6.4	17.1

Source: World Bank (2001, 1998)

By using the three DHS for the same year of adjacent countries, we are additionally able to examine to what extent there are differences in the socioeconomic determinants of undernutrition in the three countries. By applying the spatial analysis to the joint analysis of the three countries, we are additionally able to tell whether the spatial determinants cross the boundaries between the three countries or are quite distinct which would also give us a sense on the relative importance of policies versus geographic factors in causing undernutrition.

Undernutrition among children is usually measured by determining the anthropometric status of the child with most research focusing on children below six years of age. Researchers distinguish between three types of undernutrition: wasting or insufficient weight for height indicating acute undernutrition; stunting or insufficient height for age indicating chronic undernutrition; and underweight or insufficient weight for age which could be a result of both stunting and wasting. Wasting, stunting, and underweight for a child i are typically determined using a Z-score which is defined as:

$$Z_i = \frac{AI_i - MAI}{\sigma}$$

where AI refers to the individual anthropometric indicator (e.g. height at a certain age), MAI refers to the median of a reference population, and σ refers to the standard deviation of the reference population. The reference standard typically used for the calculation is the NCHS-CDC Growth Standard that has been recommended for international use by WHO (WHO, 1983; 1995).

[1] In the case of Tanzania, growth refers only to 1988-1992

The percentage of children whom Z-scores are below minus -2 standard deviations (SD) from the median of the reference category are considered as undernourished (stunted, wasted, and underweight, depending on the indicator chosen), while those with Z-Scores below -3 are considered severely undernourished. In this paper we focus on stunting, but use the Z-Score (in a standardized form) as a continuous variable to use the maximum amount of information available in the data set.

When modelling the determinants of undernutrition, one can distinguish between immediate, intermediate, and underlying determinants (see UNICEF, 1998). While undernutrition is always immediately related to either insufficient nutrient intake or the inability of the body to absorb nutrients (primarily due to illness), these are themselves caused food security, care practises, and the health environment at the household level, which themselves are influenced by the socioeconomic and demographic situation of households and communities (UNICEF; 1998; Smith and Haddad, 1999, Klasen, 1999). In order to capture this complex chain of causation, researchers have either focused on a particular level of causality (e.g. Smith and Haddad, 1999; Moradi, 1999, Pelletier, 1999), have estimated structural equations that address the interactions (e.g. Guilkey and Riphahn, 1998), have used graphical chain models to assess the causal pathways (Caputo, et al. 2002), or have used multi-level modelling techniques (e.g. Nyovani et al. 1999). With the data available, it is not always clear to separate intermediate from underlying determinants. For example, mother's education might be influencing care practises, an intermediate determinant, and the resources available to the household, an underlying determinant.

Given these difficulties, we estimate reduced form equations that mainly model factors that are mostly underlying determinants of undernutrition, although some might also be considered intermediate determinants. The most important covariates included are measures of household resources (including access to electricity and radio), access to water and sanitation, mother's education and employment status, mother's BMI as an indicator of the nutritional situation of the household, household size, the child's age and sex, and the location (urban, rural) of the household.

In previous studies on child undernutrition in Sub-saharan Africa, the influence of some of these factors has been assumed to be linear on undernutrition and we reproduce such an estimation below. However, in practice, some of these factors are likely to have non-linear effects on undernutrition.

In particular, the nutritional situation of the mother, measured using the Body Mass Index (BMI, defined as the weight in kg divided by the square of height in meters) might presumed to follow an inverse U-shape (see also Smith et al. 2001). Mothers who exhibit a very low BMI, indicating their poor nourishment, are likely to have poorly nourished children. At the same time, parents with a very high BMI might also have poorly nourished children as the obesity associated with their high BMI indicates poor quality of nutrition and might therefore indicate poor quality of nutrition for their children. Moreover, the development of undernutrition typically follows a pattern that is closely related to the age of the child. While some children are already born undernourished due to growth retardation *in utero*, the anthropometric status of children worsens considerably only after 4-6 months, when children are weaned and solid foods are introduced (WHO, 1995; Stephenson, 1999). Initially, the worsening anthropometric status shows up as acute undernutrition. But then stunting develops and worsens until about age 2-3. At that time, the body has, through reduced growth, adjusted to reduced nutritional intake and now needs fewer nutrients to maintain this smaller stature (WHO, 1995; Moradi and Klasen, 2000).

Similarly, spatial analyses of undernutrition often are confined to using region-specific dummy variables to capture the spatial dimension. After reproducing such a simple framework, we will then explore regional patterns of childhood undernutrition and, possibly nonlinear, effects of other factors within a simultaneous, coherent regression framework using a semi-parametric mixed model. Because the predictor contains usual linear terms, nonlinear effects of metrical covariates and geographic effects in additive form, such models are also called geo-additive models. Kammann and Wand (2001) propose this type of models within an empirical Bayes approach. Here, we apply a fully Bayesian approach as suggested in Fahrmeir and Lang (2001) which is based on Markov Random Field priors and uses MCMC techniques for inference and model checking.

Figure 1.1 Mean Z-score of stunting.

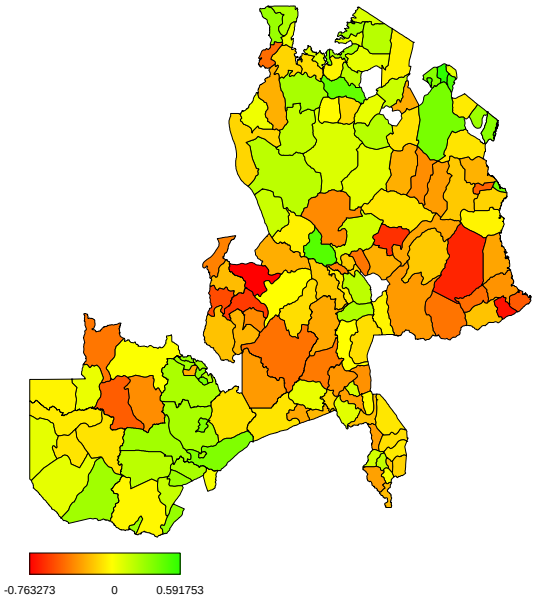
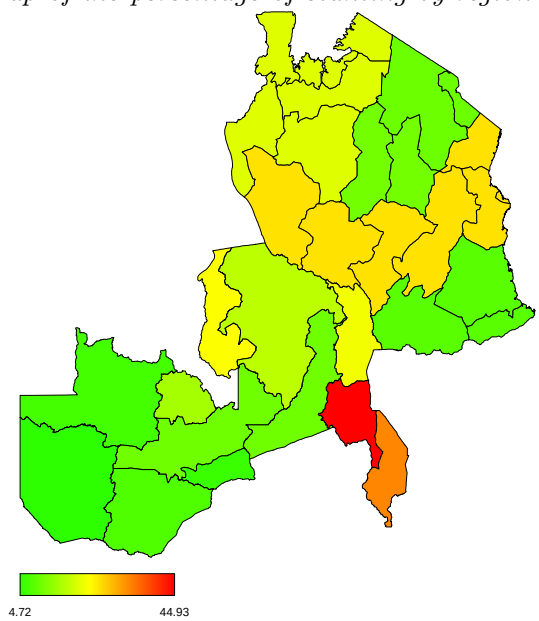


Figure 1.1 shows the small-scale, district-specific regional distribution of the (standardized) Z-scores for stunting and Figure 1.2 shows the percentage of stunting by regions (clearly reproduced in Table 1.2).

Table 1.2 Percentage of stunting by region (DHS 1992).

Regions	Percent
Malawi	
North	21.74
Central	44.93
South	33.33
Tanzania	
Coastal	25.38
Northern Highlands	10.00
Lake	20.00
Central	10.77
Southern Highlands	25.38
South	8.46
Zambia	
Central	10.38
Copperbelt	15.09
Eastern	10.38
Luapula	22.64
Lusaka	5.66
Northern	16.98
North-Western	6.60
Southern	7.55
Western	4.72

Figure 1.2 Map of the percentage of stunting by region (DHS 1992).



Obviously, there are distinct regional differences. In addition to local small-area variability, there might also be an underlying smooth spatial component which crosses borders. Figure 1.2 suggests that there are significant variations in term of stunting prevalence in the three countries and within regions in each country.

For an adequate analysis we need a flexible regression model, which can separate a smooth global spatial pattern from small-scale regional variability and simultaneously controls for demographic and socio-economic factors. In Section 4 we will compare maps obtained from such a geo-additive model with the map in Figure 1.1.

2 Semi-parametric Bayesian regression models

2.1 Observation models

Consider regression situations, where observations (y_i, x_i, w_i), $i = 1, \ldots, n$, on a metrical response y, a vector $x = (x_1, \ldots, x_p)$ of metrical covariates

and a vector $w = (w_1, \ldots, w_r)$ of categorical covariates are given. We assume that y_i given the covariates and unknown parameters are independent and Gaussian with mean η_i and a common variance σ^2 across subjects, i.e. $y_i \sim N(\eta_i, \sigma^2)$. In our application on childhood undernutrition the response is stunting measured as a standardized Z-score. Traditionally, the effect of the covariates on the response is modelled by a linear predictor

$$\eta_i = x_i'\beta + w_i'\gamma. \tag{1}$$

In such an analysis, spatial structure can be included using regional dummy variables (see below). Using smaller spatial units such as districts would in this case entail more than 200 dummy variables which would significantly reduce the degrees of freedom and could in any case not assess spatial interdependence. In this paper particular emphasis is on the nonlinear effects of the two metrical covariates "age of the child" AGC and the "mother's body mass index" BMI and, in particular, on the spatial covariate "child's district of residence". Thus, we replace the strictly linear predictor (1) by the more flexible semiparametric predictor

$$\eta_i = f_1(x_{i1}) + \cdots + f_p(x_{ip}) + f_{spat}(s_i) + w_i'\gamma. \tag{2}$$

Here, $f_1, \ldots, f_p$ are non-linear smooth effects of the metrical covariates and f_{spat} is the effect of the spatial covariate $s_i \in 1, \ldots, S$ labelling the districts in the three countries. Regression models with predictors as in (2) are sometimes referred to as geo-additive models. In a further step we may split up the spatial effect f_{spat} into a spatially correlated (structured) and uncorrelated (unstructured) effect

$$f_{spat}(s_i) = f_{str}(s_i) + f_{unstr}(s_i) \tag{3}$$

A rationale is that a spatial effect is usually a surrogate of many unobserved influences, some of them may obey a strong spatial structure and others may be present only locally. By estimating a structured and an unstructured effect we attempt to separate these effects. As a side effect we are able to assess to some extent the amount of spatial dependence in the data by observing which of the two effects is larger. If the unstructured effect exceeds the structured effect, the spatial dependence is smaller and vice versa. Such models are common in spatial epidemiology, see e.g. Besag et al. (1991).

A further extension of the predictor (2) leads to varying coefficient mixed model (VCMM)

$$\eta_i = f_1(x_{i1})z_{i1} + \cdots + f_p(x_{ip})z_{ip} + f_{str}(s_i) + f_{unst}(s_i) + w_i'\gamma. \qquad (4)$$

In varying coefficients models, it is assumed that the effect of a particular covariate z_i is not fixed but varies smoothly over the domain of a second covariate x_i. Thus, variable x_i is called the effect modifier of z_i and is usually assumed to be metrical.

2.2 Prior assumptions

In a Bayesian approach unknown functions f_j and parameters γ as well as the variance parameter σ^2 are considered as random variables and have to be supplemented with appropriate prior assumptions. In the absence of any prior knowledge we assume independent diffuse priors $\gamma_j \propto const$, $j = 1, \ldots, r$ for the parameters of fixed effects. Another common choice are highly dispersed Gaussian priors.

Several alternatives are available for the priors of the unknown (smooth) functions f_j, $j = 1, \ldots, p$. For the moment we may distinguish roughly two main approaches for Bayesian semiparametric modelling. These are basis functions approaches with adaptive knot selection and approaches based on smoothness priors. In the following we will focus on the latter one. Several alternatives have been proposed for specifying a smoothness prior for the effect f of a metrical covariate x. Among others, these are random walk priors (Fahrmeir and Lang, 2001), Bayesian smoothing splines (Hastie and Tibshirani, 2000) and Bayesian P-splines. In this paper we focus on random walk priors. We also compared our results with Bayesian smoothing splines and P-splines but the estimated functions were more or less undistinguishable.

For the random walk prior, let us first consider the case of a metrical covariate x with *equally spaced observations* x_i, $i = 1, \ldots, m$, $m \leq n$. Suppose that $x_{(1)} < \ldots < x_{(t)} < \ldots < x_{(m)}$ is the ordered sequence of distinct covariate values. Define $f(t) := f(x_{(t)})$ and let $f = (f(1), \ldots, f(t), \ldots, f(m))'$ denote the vector of function evaluations. Then a first or second order random walk prior for f is defined by

$$f(t) = f(t-1) + u(t) \quad \text{or} \quad f(t) = 2f(t-1) - f(t-2) + u(t) \qquad (5)$$

with Gaussian errors $u(t) \sim N(0; \tau^2)$ and diffuse priors $f(1) \propto$ const, or $f(1)$ and $f(2) \propto$ const, for initial values, respectively. A first order random walk penalizes abrupt jumps $f(t) - f(t-1)$ between successive states and a second order random walk penalizes deviations from the linear trend $2f(t-1) - f(t-2)$. Random walk priors may be equivalently defined in a more symmetric form by specifying the conditional distributions of function evaluations $f(t)$ given its left *and* right neighbors, e.g. $f(t-1)$ and $f(t+1)$ in the case of a first order random walk. Thus, random walk priors may be interpreted in terms of locally polynomial fits. A first order random walk corresponds to a locally linear and a second order random walk to a locally quadratic fit to the nearest neighbors. Of course, higher order autoregressions are possible but practical experience shows that the differences in results are negligible. For the case of *nonequally spaced observations* random walk priors must be modified to account for nonequal distances $\delta_t = x_{(t)} - x_{(t-1)}$ between observations. Random walks of first order are now specified by

$$f(t) = f(t-1) + u(t), \quad u(t) \sim N(0; \delta_t \tau^2), \tag{6}$$

i. e., by adjusting error variances from τ^2 to $\delta_t \tau^2$. Random walks of second order are defined by

$$f(t) = \left(1 + \frac{\delta_t}{\delta_{t-1}}\right) f(t-1) - \frac{\delta_t}{\delta_{t-1}} f(t-2) + u(t), \tag{7}$$

$u(t) \sim N(0; w_t \tau^2)$, where w_t is an appropriate weight. Several possibilities are conceivable for the weights, see Fahrmeir and Lang (2001) for a discussion. However, in this analysis, we use a second random walk prior for metrical covariates.

The trade off between flexibility and smoothness of f is controlled by the variance parameter τ^2. In our approach we want to estimate the variance parameter and the smooth function simultaneously. This is achieved by introducing an additional hyperprior for τ^2 in a further stage of the hierarchy. We choose a highly dispersed but proper inverse gamma prior $p(\tau^2) \sim IG(a; b)$ with $a = 1$ and $b = 0.005$. In analogy, we also define for the overall variance σ^2 a highly dispersed inverse gamma prior.

Let us now turn our attention to the spatial effects f_{str} and f_{unstr}. For the spatially correlated effect $f_{str}(s)$, $s = 1, \ldots, S$, we choose Markov random

field priors common in spatial statistics (Besag, *et al.* 1991). These priors reflect spatial neighborhood relationships. For geographical data one usually assumes that two sites or regions s and r are neighbors if they share a common boundary. Then a spatial extension of random walk models leads to the conditional, spatially autoregressive specification

$$f_{str}(s) \mid f_{str}(r), r \neq s, \tau^2 \sim N\left(\sum_{r \in \partial_s} f_{str}(r)/N_s, \tau^2/N_s\right) \tag{8}$$

where N_s is the number of adjacent regions, and

$r \in \partial_s$ denotes that the region r is a neighbor of region s. Thus the conditional mean of $f_{str}(s)$ is an unweighed average of function evaluations for neighboring regions. Again the variance τ^2_{str} controls the degree of smoothness.

For a spatially uncorrelatated (unstructured) effect f_{unstr} common assumptions are that the parameters $f_{unstr}(s)$, are i.i.d. Gaussian

$$f_{unstr}(s) \mid \tau^2{}_{unstr} \sim N(0, \tau^2_{unstr}) \tag{9}$$

Also here, variance or smoothness parameters $\tau^2_j, j = 1, \ldots, p, str, unstr$, are also considered as unknown and estimated simultaneously with corresponding unknown functions f_j. Therefore, hyperpriors are assigned to them in a second stage of the hierarchy by highly dispersed inverse gamma distributions $p(\tau^2_j) \sim IG(a_j, b_j)$ with known hyperparameters a_j and b_j.

2.3 Posterior inference

Bayesian inference is based on the posterior and is carried out using recent MCMC simulation techniques. For the predictor (4), let α denote the vector of all unknown parameters in the model. Then, under usual conditional independence assumptions, the posterior is given by

$$p(\alpha|y) \propto \prod_{i=1}^{n} L_i(y_i, \eta_i) \prod_{j=1}^{p} \{p(f_j|\tau^2_j)p(\tau^2_j)\} p(f_{str}|\tau^2_{str}) p(f_{unstr}|\tau^2_{unstr}) \prod_{j=1}^{r} p(\gamma_j) p(\sigma^2),$$

where f_j, $j = 1, \ldots, p$, are vectors of function evaluation corresponding to the functions f_j. The full conditionals for the parameter vectors $f_1, \ldots, f_p$ as well as the full conditionals for f_{str}, f_{unstr} and fixed effects parameters γ are multivariate Gaussian. For the variance components τ_j^2, $j = 1, \ldots, p, str, unstr$ and σ^2 the full conditionals are inverse gamma distributions. Thus, Gibbs sampler can be used for MCMC simulation, with successive draw from the full conditionals for

$f_1, \ldots, f_p, f_{str}, f_{unstr}, \tau_j^2$, $j = 1, \ldots, p$ and σ^2. Sampling efficiency from the Gaussian full conditionals of non-linear functions is guaranteed by Cholesky decompositions for band matrices.

2.4 Bayesian measures of model complexity and fit

From the Bayesian perspective, model comparison is done by trading off the measure of fit, typically a deviance statistics, and a measure of complexity, the number of free parameters in the models. Since increasing complexity is accompanied by better fit, proposals are formally based on minimizing a measure of expected loss on a future replicate data set (For more details, see, Spiegelhalter *et. al.*, 2001; Efron 1996). In hierarchical surveys dataset parameters may outnumbers observations. We adopt the measure of complexity and fit suggested by Spiegelhalter *et. al.*, (2001). They used an information-theoretic argument to suggest the Deviance Information Criterion (DIC) as a measure of fit and model complexity. It penalizes the posterior mean deviance $\bar{D}$, which is a measure of fit, by a complexity measure pD for the effective number of parameters in a model, as the difference between the posterior mean of the deviance and the deviance at the posterior estimates of the parameters of interest. pD is shown to be approximately the trace of the product of Fisher's information and the posterior covariance matrix, which can be obtained easily from a Markov chain Monte Carlo (MCMC) analysis. They also argued that, for normal models, pD corresponds to the trace of the 'hat' matrix projection observations onto fitted values.

Let $f(y)$ be some fully specified standardizing term that is a function of the data alone, then

$$pD = \bar{D} - D(\bar{\theta}) \tag{10}$$

where $D(\theta) = -2logp(y|\theta) + 2logf(y)$, the Bayesian deviance.

The *Deviance Information Criterion* (DIC), defined as a 'plug-in' estimate of fit, plus twice the effective number of parameters, is defined as

$$DIC = D(\bar{\theta}) + 2pD = \bar{D} + pD. \tag{11}$$

For more details see, Spiegelhalter *et. al.*, (2001). Thus, the posterior mean of the deviance is penalized by the effective number of model parameters pD. The models in chapter 3 can be validated by analyzing the DIC, which decreases for models with covariates of high explanatory value.

2.5 Coding of Categorical covariates

Categorical covariates such as the child's gender, the educational achievement of the respondent, the household size, the income of the family and socioeconomic covariates are effect coded.

The effect coding for this matter is preferred because of ease of interpretation and its advantage over the dummy coding in computing the reference category. An effect coding variable is defined by,

$$x^{(j)} = \begin{cases} 1 & \text{if category } j \text{ is observed} \\ -1 & \text{if category } k \text{ is observed} \\ 0 & \text{else} \end{cases} \qquad j = 1...q.$$

In this coding the reference category k is given by the vector (-1,...,-1). The effect of categorical covariates (demographic and socio-economic variables) are considered as fixed and constant and are estimated jointly with metrical and spatial covariates.

3 Data and results

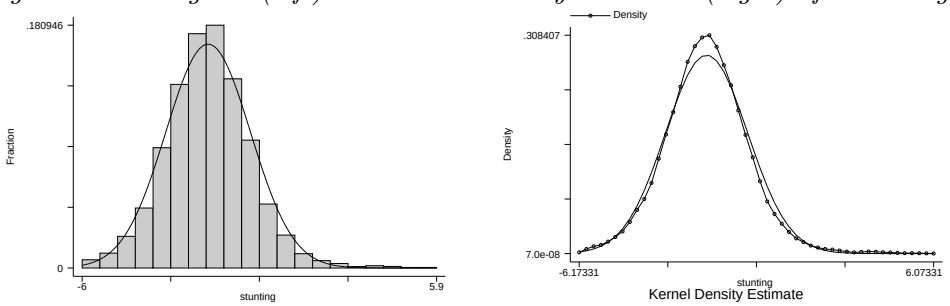
3.1 Data, Descriptive Statistics, and Models

The data used are from the 1992 Demographic and Health Surveys for the three countries. The DHS collect information on a nationally representative

sample of women in child-bearing age (15-49). The questionnaire collects socioeconomic indicators for the respondent and her partner as well as the household she resides in, and then gathers a large amount of information on fertility patterns, health and care practises, health knowledge, and assesses the anthropometric status of all children of these women who were born within the past five years. Unfortunately, the surveys do not generate an income variable and we therefore rely on household assets as a proxy for the income situation of the households which has been found to be quite reliable by Filmer (1999).

The 1992 DHS data sets of Malawi, Tanzania and Zambia are pooled together to form one data set with the same socio-economic, demographic and health characteristics of the household. This is possible because the DHS surveys are carried out in standardized form, with the same list of socio-economic and demographic characteristics. The sample now comprises a total of 844 clusters in 156 districts in the three countries. We take the clusters located in a particular district as representative of that district.

Figure 3.1 Histogram (left) and kernel density estimates (right) of "stunting".



The geographical distribution of the standardized Z-scores for the response variable stunting, averaged within districts, was already displayed in Figure 1.1. Figure 3.1 shows a histogram and kernel density estimates of the distribution of the Z-scores, together with a normal density, with mean and variance estimated from the sample. This gives clear evidence that a Gaussian model is a reasonable choice for inference.

Empirical distributions of categorical covariates, together with codings used in the analysis, are given in Table 3.2. Other categorical covariates, such as the employment situation of the mother, household size and type of toilet facility, turned out to be non-significant in the preliminary data analysis and

were thus omitted . While all three countries do relatively poorly on the reported socioeconomic indicators, there are significant differences between the countries as well. In particular, households in Zambia appear to be better off in terms of access to electricity, radio, and female educational attainment, which was already apparent from Table 1.1 which showed that income and education levels were higher in Zambia. This country is also more heavily urbanized than the other two. Malawi and Tanzania are more similar, with Malawi doing somewhat worse on access to electricity. Malawi also has worse educational attainment at the lower levels but slightly higher among the highest levels than Tanzania.

The empirical distributions of the metrical covariates mother's body mass index (BMI) and child's age are shown in Figures 3.2. Note that only children not older than five years are included in the sample.

Table 3.2 Factors analyzed in childhood undernutrition studies

Factor	Malawi (%)	Tanzania(%)	Zambia (%)	coding
Residence				
Urban	25.5%	15.7%	42.7%	1 :urban
Rural	74.5%	84.3%	57.3%	-1 :rural, ref. cat.
Has radio				
No	54.8%	64.2%	57.2%	-1: ref. cat.
Yes	45.0%	34.1%	42.4%	1: yes
Has electricity				
No	94.8%	92.8%	80.5%	-1: ref. cat.
Yes	5.0%	5.6%	19.2%	1: yes
Educational attainment				
No educ.	41.6%	37.2%	17.9%	-1 : ref. cat.(incl. inc. prim.)
Incomp. prim.	42.8%	18.9%	30.3%	
Compl. prim.	10.1%	40.6%	32.8%	cat. 1 (incl. inc. sec.)
Incomp. sec.	3.5%	3.0%	15.3%	
Compl. sec.	1.7%	0.1%	2.1 %	cat. 2 (incl. higher educ.)
Higher	0.2%	0.2%	1.5%	
Sex of child				
Male	50.7%	50.1%	50.1%	1: male
Female	49.3%	49.9	49.9%	-1: ref. cat.
Mean BMI	21.96	21.75	21.96	metrical
District	spatial covariate			

We analyzed several models, differing in complexity. The simple linear model

$$\eta_i = \alpha + w_i'\gamma. \tag{12}$$

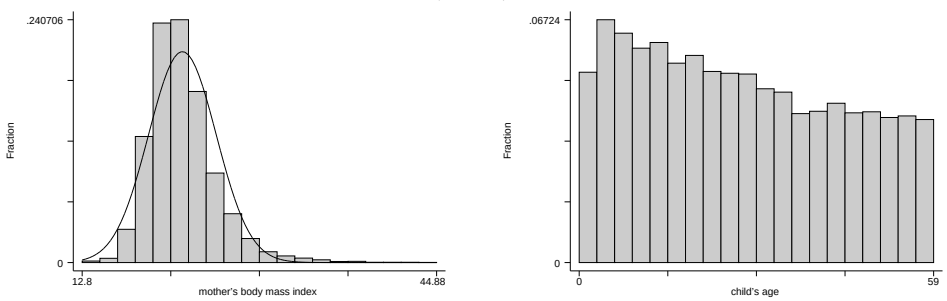
assumes that the fixed effects of covariates are the same for all three countries, and includes dummies for regions.

Based on previous analysis carried out separately for each country (Kandala *et al.*, 2001), we choose a geo-additive model with interactions between country-effects and educational attainment as well as the availability of electricity in a next step. Taking Tanzania as the reference country, we define 0/1-dummies ZA and MA for Zambia and Malawi, and arrive at the model

$$\begin{aligned}\eta &= \alpha + f_1(agc_i) + f_2(bmi_i) + f_{str}(s_i) + f_{unstr}(s_i) \\ &+\beta_1 edu1_i + \beta_2 edu2_i + \beta_5 edu1dma_i + \beta_6 edu2dma_i + \\ &\beta_7 edu1dza_i + \beta_8 edu2dza_i + \beta_9 elcdma_i + \beta_{10} elcdza_i + w_i'\gamma, \end{aligned} \tag{3.2}$$

This geo-additive model assumes that the nonlinear effects $f_1, \ldots$ and the fixed effects γ are the same for all three countries. This was confirmed by prior separate analyses of the non-linear effects in each of the countries which were found to be remarkably similar. Moreover, in a further step, we also analyzed a varying coefficient model (see equation 2.4), to see if the patterns of the nonlinear effects for mother's body mass index and child's age differ between countries. It turned out, however, that there is no significant difference, so that the effects of BMI and child's age on stunting can be assumed to follow the same general pattern in all three countries.

Figure 3.2 Histograms of "mother's body mass index" (left) and child's age (right).



3.2 Results

The estimates of fixed effects of the covariates in w of the linear model (see equation 3.1) are given in Table A1, and the linear effects of BMI and child's age are also shown in Figure A1. The regional-fixed effects are also shown in the maps of Figures A2.

The linear model assumes a positive relationship between mother's BMI and stunting and a negative relationship between the child's age and stunting. As we show below, this glosses over important non-linearities in the effects.

The other fixed effects are mostly as expected. Children from educated mothers in urban areas with access to electricity and a radio are better nourished. Female children are also slightly less stunted which has also been found in other studies (Svedberg, 1996; Klasen, 1996; Hill and Upchurch 1995). The regional fixed effects are mostly quite significant suggesting that the socioeconomic variables are unable to account for a considerable portion of this regional variation.

Table A2 contains the fixed effects for the model 3.2, and the non-linear effects of BMI and child's age are shown in Figure A1. In the left-hand map of Figure A3 we show the mean Z-scores by district based on the socioeconomic covariates; in the right-hand map we then subtract the predicted Z-score from the left-hand figure from the raw Z-score from figure 1.1 to get the raw spatial residual, i.e. the component of the Z-score not explained by the socioeconomic variables. This is then allocated to structured and unstructured effects. The posterior mean estimates of the structured smooth spatial component f_{str} and the unstructured random component f_{unst} are shown in the maps of Figures A4 and A5.

In addition, posteriori probability maps indicate significance of the spatial effects (white/black = significantly positive/negative effect on the Z-score, grey = nonsignificant).

Note that the spatial effects are centered about zero, i.e. the average Z-score over all districts is zero, while the overall level is estimated through the intercept term α. Table A3 also shows averages of the spatial effects computed separately for the three countries, indicating differences in the overall level between countries. Note that these country effects are thus recovered from the spatial analysis based on the estimated average district-effect in each country and are not separately estimated.

Before commenting on the substantive results, it is important to point out that the fit criteria improve considerably in this model vis-a-vis the previous one. Also the DIC, which penalizes for the additional parameters is much improved suggesting a considerably better fit of this flexible model.

The fixed effects in Table A2 are virtually identical to the linear model. The only country interaction that turned out to be significant and thus were

retained in the model were the interactions with mother's education and electricity. Here we find that the positive effect of high mother's education is much smaller in Zambia and Malawi than in Tanzania. Similarly, the positive effect of having access to electricity is also significantly smaller in Malawi and Zambia. It thus appears that these two socioeconomic indicators have a much larger effect in Tanzania than elsewhere.

The recovered country effects from the district analysis are also reported here and suggest that nutrition in Tanzania is, after controlling for socioeconomic effects, significantly better than in the other two countries, although the effect is small.

The left panel of Figure A1 shows the flexible modelling of the effect of the BMI of the mother. Shown are the posterior means together with 80 % pointwise credible intervals. We find the influence to be in the form of an inverse U shape. While the inverse U looks nearly symmetric, the descending portion exhibits a much larger range in the credible region. This appears quite reasonable as obesity of the mother (possibly due to a poor quality diet) is likely to pose less of a risk for the nutritional status of the child as very low BMIs which suggest acute undernutrition of the mother. The Z-score is highest (and thus stunting lowest) at a BMI of around 30-35.

Clearly, this inverse U has not been picked up in the linear fit and also a simple polynomial would not pick up the differences between the ascending and descending portion. The right panel of Figure A1 shows the effect of the child's age on its nutritional status. As suggested by the nutritional literature, we are able to discern the continuous worsening of the nutritional status up until about 20 months of age. This deterioration set in right after birth and continues, more or less linearly, until 20 months. Such an immediate deterioration in nutritional status is not quite as expected as the literature typically suggests that the worsening is associated with weaning at around 4-6 months. One reason for this unexpected finding could be that, according to the surveys, most parents give their children liquids other than breastmilk shortly after birth which might contribute to infections.

After 20 months, stunting stabilizes at a low level. Through reduced growth and the waning impact of infections, children are apparently able to reach a low-level equilibrium that allows their nutritional status to stabilize.

We also see a blip around 24 months of age. This is picking up the effect of a change in the data set that makes up the reference standard. Until 24 months,

the currently used international reference standard is based on white children in the US of high socioeconomic status, while after 24 months, it is based on a representative sample of all US children (WHO, 1995). Since the latter sample exhibits worse nutritional status, comparing the Tanzanian children to that sample leads to a sudden improvement of their nutritional status at 24 months. This drawback of the reference standard is one reason for WHO's current efforts to construct a new reference standard (WHO, 1999).

Figure A3 (left) shows that the socioeconomic effects are able to explain a fair amount of the spatial variation of undernutrition in the three countries. This can also be seen that the range of standardized Z-scores in the right-hand figure (which shows the spatial residual) is only about half as large as the total variation was in Figure 1.1.

But the spatial residuals in the right-hand side of figure A3 show that much of the variation in stunting remains to be explained. Moreover, one can see already that the spatial residuals transcend the borders. While there is some clear demarcation between the better districts in Western Tanzania and the worse districts just across the border in Zambia, there appears to be a continuum of negative spatial residuals that runs from Northeastern Zambia, Northern and Central Malawi, and into Southern Tanzania.

These spatial effects are then allocated by the model into structured and unstructured effects which are shown in Figures A4 and A5. Several important findings emerge. First, many of these structured spatial effects are significant. Thus we clearly have a pattern of worse nutrition in Eastern and Northeastern Zambia, Central Malawi, and Southern Tanzania. Conversely, Z-scores are significantly better in Northern Tanzania. Second, while these structured effects suggests worse undernutrition in a belt ranging from Northern Zambia to Southern Tanzania, it is interesting to note that the districts in Northern Malawi, and South-Western Tanzania are not significant components in that belt. Thus while some spatial residuals do spill significantly across borders, e.g. between Northern Zambia and Central Malawi , some borders do seem to matter in the sense that spatial residuals remain noticeably distinct in the analysis on the two sides of borders.

Third, the structured effects are clearly more important than the unstructured random effects. Only very few of the unstructured effects are significant and the range of unstructured effects is much smaller than the structured ones. Thus most of the spatial residual effect was allocated to the structured ones, i.e. the effects where neighborhood matters. This seems reasonable given

the strong spatial pattern that one can determine in the total spatial residual (Figure A3 right). [2]

The few unstructured effects that do exist are interesting. First, in Tanzania, we find that stunting in the capital Dar es Salaam is significantly better after accounting for socioeconomics and structured spatial effects. This situation of better undernutrition in large cities is not replicated in Zambia. In fact, Lusaka has no better stunting, even though surrounding districts have significantly better undernutrition rates; some cities in the copperbelt are actually doing significantly worse (see the small districts at Zambia's Northern border in the central part of the country). This may be related to the effect of the decline in copper production and the impact of general economic decline and structural adjustment policies that have affected urban areas more than rural areas (World Bank, 2000). Moreover, there is one district in Northern Malawi and two in Northern Tanzania that have significantly positive unstructured effects.

These subtle effects are clearly not captured by the provincial fixed effects in Figure A2 where both the spatial pattern as well as the fine differentiations within provinces are not adequately captured.

The clear structured pattern begs for an explanation. None of the socioeconomic variables we tried in addition to the ones mentioned are able to reduce these pronounced spatial effects. One common factor to most of the areas that are negatively affected are that these areas are at comparatively low elevations while the areas of positive spatial effects tend to be at higher elevations. This distinction is most noticeable and clear in the South-North divide in Tanzania, but also noticeable elsewhere. The difference could well be due to differences in disease prevalence such as Malaria, Schistosomasis, and other diseases that thrive at lower elevations and are particularly problematic along the Rift Valley. In an exploratory analysis, we compared the spatial pattern of prevalence of fever, diarrhea, cough or any of the three illnesses combined with the structured spatial pattern and found that the spatial distribution of fever (presumably related to Malaria) has a fairly close resemblance to the structured spatial effects while the others do not

[2] The allocation into structured and unstructured effects appears quite robust as it did not change greatly when the three countries were considered separately (Kandala, 2002), and it also is reasonable in the sense that district with positive spatial residuals that are surrounded by districts with negative spatial residuals will influence each other and average out the effects with the residual of that process being picked up by the unstructured effect.

appear to play a significant role. Future work should explore this linkage further [3] Moreover, the poor nutritional status in Northeastern Zambia could additionally be related to the poor access to health facilities and the general remoteness of these areas which are poorly served with transportation links (World Bank, 2000). These issues deserve closer attention and this procedure is merely able to highlight the important spatial patterns of undernutrition without being able to fully explain them.

Quite clearly, the methods used here are able to identify more subtle socioeconomic and spatial influences on undernutrition than reliance on linear models with regional dummy variables would have allowed. As such, they are useful for diagnostic purposes to identify the need to find additional variables that can account for this spatial structure. Moreover, even if the causes of the spatial structure are not fully explained, one can use this spatial information for poverty mapping and associated targeting purposes, which is gaining increasing importance in policy circles that attempt to focus the allocation of public resources to the most deprived sections of the population.

4 Conclusion

In this paper we pooled the data from the 1992 Demographic and Health surveys of Malawi, Tanzania, and Zambia to model the socioeconomic and spatial determinants of undernutrition. We found strong support for our approach of flexibly modelling some covariates that clearly have non-linear influences and for including a spatial analysis. The spatial analysis shows distinct spatial patterns that point to the influence of omitted variables with a strong spatial structure or possibly epidemiological processes that account for this spatial structure.

The maps generated could be used for targeting development efforts at a glance, or for exploring relationships between welfare indicators and others variables. For example, a mortality or undernutrition map could be overlaid with maps of other types of data, say on poverty, agro-climatic or other environmental characteristics. The visual nature of the maps may highlight

[3] The measure of disease prevalence used here, recall of whether anyone in the household had been ill with fever, cough, or diarrhea in the past 2 weeks is less perfect as it is quite subjective, based on a short-term recall, and has considerable noise. Future work needs to address the question of disease environment more closely

unexpected relationships that would be overlooked in a standard regression analysis.

Acknowledgement: This research was supported by the Deutsche Forschungsgemeinschaft, Sonderforschungsbereich 386 "Statistische Analyse diskreter Strukturen". We thank participants at seminars in Munich and at IFPRI in Washington DC for helpful comments and discussions. We also want to thank Macro Intl. for providing us with the district locations of the clusters in the three DHS data sets.

Table A1 Linear Model (Model 3.1)

Deviance: 12266.3 DIC :12289.1

Variable	mean	10% quant.	90%quant.
Constant	0.08	-0.001	0.17
Mother's BMI	0.02	0.02	0.03
Child's age	-0.02	-0.02	-0.02
Urban	0.08	0.07	0.10
Rural	-0.08	-0.10	-0.07
Male	-0.05	-0.06	-0.04
Female	0.05	0.04	0.06
No edu. and incompl. prim. edu. Tan.	-0.13	-0.16	-0.11
Compl.primary edu. and incompl. sec. edu Tan.	-0.05	-0.08	-0.02
Secondary edu. and higher Tan.	0.19	0.13	0.24
Has electricity Tan.	0.08	0.06	0.10
No electricity Tan.	-0.08	-0.10	-0.06
Has radio Tan.	0.05	0.04	0.06
No radio Tan.	-0.05	-0.06	-0.04
Coastal	0.02	-0.01	0.05
Northern Highlands	0.22	0.17	0.27
Lake	0.15	0.12	0.18
Central	0.03	-0.02	0.07
Southern Highlands	-0.05	-0.09	-0.01
South	-0.24	-0.28	-0.19
North Mal.	0.03	-0.01	0.07
Central Mal.	-0.07	-0.11	-0.03
South Mal.	-0.01	-0.024	0.05
Central	0.08	0.03	0.14
Copperbelt	0.01	-0.04	0.05
Eastern	-0.08	-0.14	-0.03
Luapula	-0.22	-0.28	-0.17
Lusaka	0.04	-0.01	0.09
Northern	-0.17	-0.22	-0.11
North-Western	-0.06	-0.13	0.01
Southern	0.17	0.13	0.22
Western	0.14	0.08	0.21

Figure A1 Effects of mother's body mass index (left) and child's age (right) on stunting for Model 3.1 and 3.2.

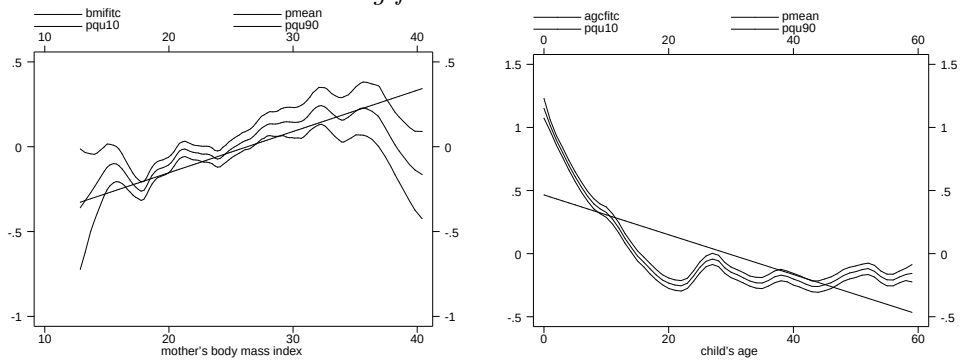


Table A2 Fixed effects for Model 3.2
Deviance: 10826.2 DIC : 10938

Variable	mean	10% quant.	90%quant.
constant	0.29	0.23	0.34
Urban	0.08	0.07	0.10
Rural	-0.08	-0.1	-0.07
Male	-0.05	-0.06	-0.04
Female	0.05	0.04	0.06
No edu. and incompl. prim. edu. Tan.	-0.21	-0.29	-0.12
Compl.primary edu. and incompl. sec. edu Tan.	-0.16	-0.24	-0.07
Secondary edu. and higher Tan.	0.37	0.20	0.53
Add. effect of no edu. and incompl. prim. edu. Mal.	0.08	-0.02	0.19
Add. effect of Compl.primary edu. and incompl. sec. edu. Mal.	0.16	0.06	0.26
Add. effect of Secondary edu. and higher Mal.	-0.24	-0.45	-0.06
Add. effect of no edu. and incompl. prim. edu. Zam.	0.07	-0.02	0.16
Add. effect of Compl.primary edu. and incompl. sec. edu. Zam.	0.10	0.01	0.19
Add. effect of Secondary edu. and higher Zam.	-0.17	-0.35	-0.001
Has electricity Tan.	0.14	0.10	0.18
No electricity Tan.	-0.14	-0.18	-0.10
Has Radio Tan.	0.05	0.04	0.06
No Radio Tan.	-0.05	-0.06	-0.04
Add. effect of electricity Mal.	-0.02	-0.07	0.04
Add. effect of no electricity Mal.	0.02	-0.04	0.07
Add. effect of electricity Zam.	-0.07	-0.12	-0.03
Add. effect of no electricity Zam.	0.07	0.03	0.12

Figure A2 provincial fixed-effects Model 3.1

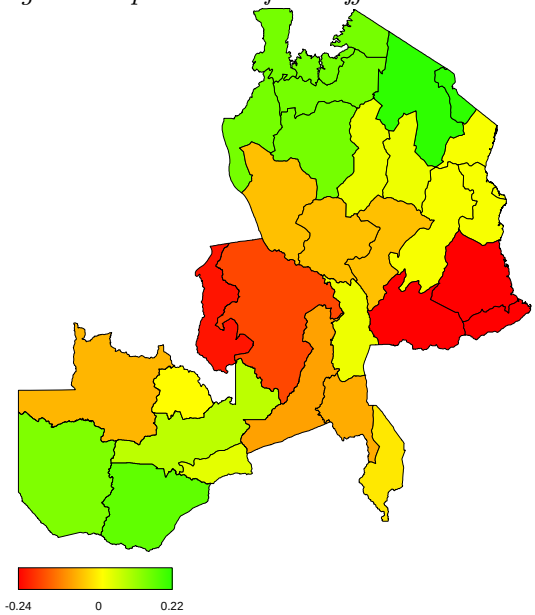


Figure A3 Mean of stunting predicted by the covariates for Model 3.2 (left) and Raw spatial residual of stunting for Model 3.2 (right).

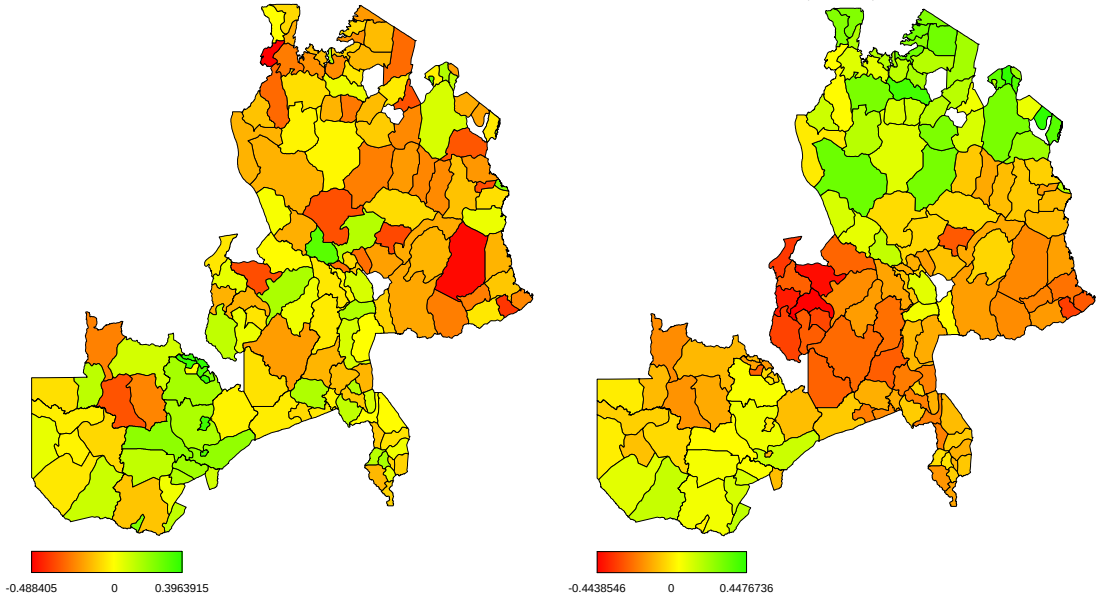


Figure A4 Structured posterior mean (left) and posterior probabilities (right) of stunting for Model 3.2

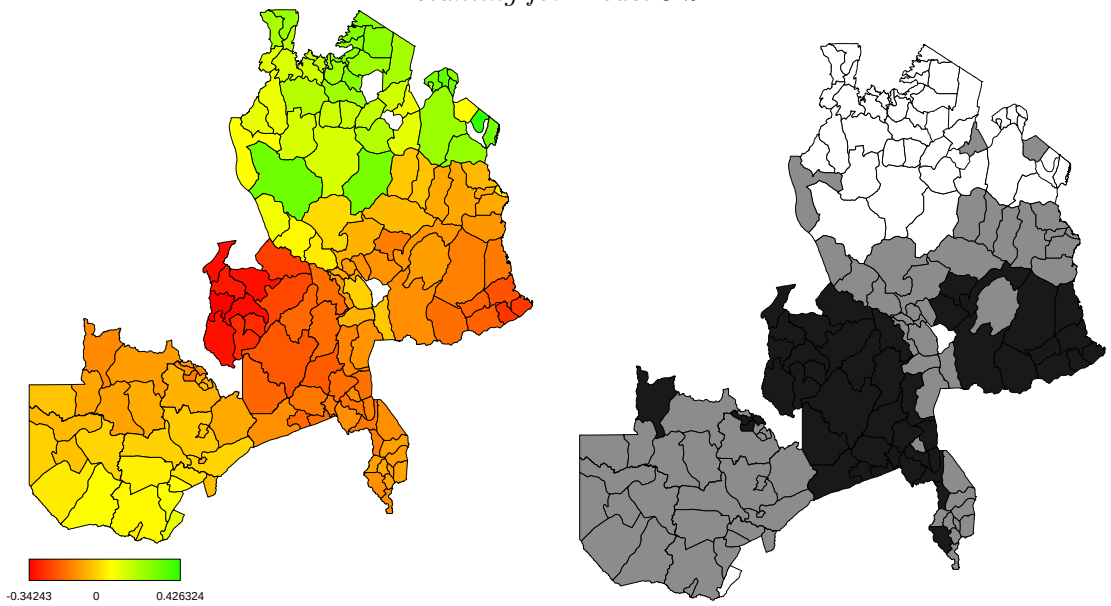


Figure A5 Unstructured posterior mean (left) and posterior probabilities (right) of stunting for Model 3.2

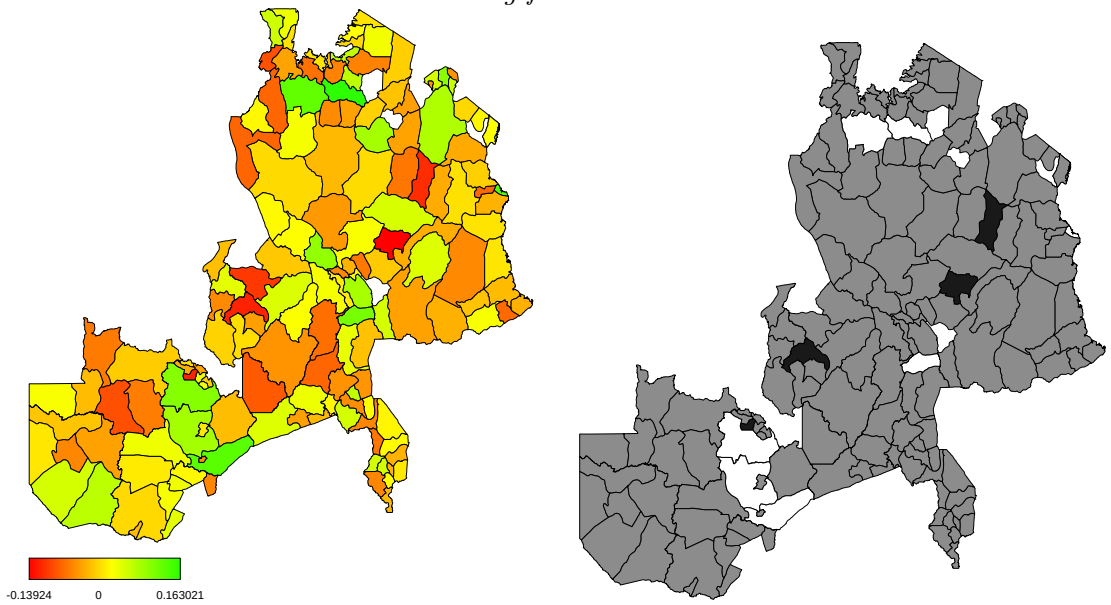


Table A3 Dummies for country main effects Model 3.2

Dummy variable	Model II
Dummy for Tanzania	0.1
Dummy for Malawi	-0.1
Dummy for Zambia	-0.1

References

[1] Besag, J., York, Y. and Mollie, A. (1991). *Bayesian Image Restoration with two Applications in Spatial Statistics (with discussion). Ann. Inst. Statist. Math.,* **43***, 1-59.*

[2] Caputo, A., R. Foraita, S. Klasen, and I.Pigeot (2002). Undernutrition in Benin: An Analysis based on Graphical Models. *Social Science and Medicine* (forthcoming).

[3] Elbers, C., Lanjouw, J.O., Lanjouw, P. (2001). *Welfare in Villages and Towns: Micro-Level Estimation of Poverty and Inequality. Tinbergen Institute Working paper No.* **2000-0029/2***, (at* `http://www.tinbergen.nl/`*.*

[4] Fahrmeir, L. and Lang, S. (2001). *Bayesian Inference for Generalized Additive Mixed Models Based on Markov Random Field Priors. Applied Statistics (JRSS C),* **50***, 201-220.*

[5] Guilkey, D. and R. Riphahn (1998). The Determinants of Child Mortality in the Philippines: Estimation of a Structural Model. *Journal of Development Economics*56: 281-305.

[6] Hastie, T. and Tibshirani, R. (1993). *Varying-coefficient Models. J. R. Statist. Soc. B,* **55***, 757-796.*

[7] Hastie, T. and Tibshirani, R. (2000). *Bayesian Backfitting. Statistical Science* **15***: 193-223.*

[8] IMF (2000). *A better world for all: progress towards the international development goals. International Monetary Fund, Organisation for Economic Co-operation and Development, United Nations, and World Bank, 2000*

[9] Kamman, E.E and Wand M. P. (2001). *Geoadditive models, Department of Biostatistics, School of Publicc Health, Havard University.*

[10] Klasen, S. (1996). *Nutrition, Health, and Mortality in Sub-Saharan Africa: Is there a Gender Bias? Journal of Development Studies* **32***: 913-932.*

[11] Klasen, S. (1999). *Malnourished and low mortality in South Asia, better nourished and dying young in Africa: What can explain this puzzle? SFB 386 Discussion Paper No. 214. University of Munich.*

[12] Lang, S. and Brezger, A. (2000a). *Bayesian P-Splines. Proc. of the 15th International Workshop on Statistical Modelling in Bilbao, 318-323.*

[13] Lang, S. and Brezger, A. (2000b). BayesX. *Software for Bayesian Inference Based on Markov Chain Monte Carlo Simulation Techniques, University of Munich.*

[14] Moradi, A. (1999). *Determinanten der Unterernhrung in Kenya, Zambia, und Indien.* Master's Thesis, University of Munich.

[15] Moradi, A. and S. Klasen (2000). *The Nutritional Status of Elites in India, Kenya, and Zambia: An Appropriate Guide for Developing International Reference Standards for Undernutrition? SFB 386 Discussion Paper No. 217. University of Munich.*

[16] N.B. Kandala (2002). *Spatial modelling of Socio-Economic and Demographic Determinants of Childhood Undernutrition and Mortality in Africa, Ph.D Thesis, University of Munich, Shaker Verlag*

[17] N.B. Kandala, S. Lang, S. Klasen, and L. Fahrmeir (2001). *Semiparametric Analysis of the Socio-Demographic Determinants of Undernutrition in Two African Countries. Research in Official Statistics, EUROSTAT, Vol. 4 No.***1***:81-100.*

[18] Nyovani, J.M., Z. Matthews, and B. Margetts (1999). *Heterogeneity of Child Nutritional Status between Households: A Comparison of six Sub-Saharan African Countries. Population Studies* **53***: 331-343.*

[19] Pelletier, D. (1998). Malnutrition, Morbidity, and Child Mortality in Developing Countries, In: United Nations (eds.) *Too Yuong too Die: Genes or Gender?* New York: United Nations.

[20] Pritchett, L. and L. Summers (1996). Wealthier is healthier. *Journal of Huamn Resources*31: 841-868.

[21] Sen, A. (1999). *Development as Freedom. Oxford: Oxford University Press.*

[22] Smith, L. and L. Haddad. (1999) *Explaining Child Malnutrition in Developing Countries.* IFPRI Research Report No. 111, Washington DC: IFPRI.

[23] Smith, L. and L. Haddad. (2001) *The Importance of Women's Status for Child Nutrition in Developing Countries.* Mimeographed, IFPRI. Washington DC: IFPRI.

[24] Smith, M. and Kohn, R. (1996). *Nonparametric regression using Bayesian variable selection. Journal of Econometrics,* **75***, 317-343.*

[25] Stephenson, C. (1999). *Burden of Infection on Growth Failure. Journal of Nutrition, Supplement, 534S-538S.*

[26] Spiegelhalter D., Best N., Carlin B., and Van der Line A. (2001). *Bayesian measures of models complexity and fit.*

[27] Svedberg, P. (1996). *Gender Bias in Sub-Saharan Africa: Reply and Further Evidence. Journal of Development Studies* **32***: 933-943.*

[28] Svedberg, P. (1999). *Poverty and Undernutrition.* New York: Oxford University Press.

[29] UN (2000). *A Better World for All.* New York: United Nations.

[30] UNICEF (1998). *The State of the World's Children. New York: UNICEF.*

[31] WHO (1983). *Measuring Change in Nutritional Status. Geneva: WHO.*

[32] WHO (1995). *Physical Status: The Use and Interpretation of Anthropometry. WHO Technical Report Series No. 854. Geneva: WHO.*

[33] WHO (1999). *Infant and Young Child Growth: The WHO Multicentre Growth Reference Study. Executive Board: Implementation of Resolutions and Decisions EB105/Inf.doc/1. Geneva: WHO.*

[34] World Bank (1995). *Zambia Poverty Assessment. Washington, DC: The World Bank.)*

[35] World Bank (1998). *Confronting AIDS.* Washington DC: The World Bank.

[36] World Bank (2000). *Profile of poverty in Malawi, 1998. Poverty analysis of the Malawian Integrated Household Survey, 1997-98. Washington, DC: The World Bank.*

[37] World Bank (2001). *World Development Indicators*. Washington DC: The World Bank.