DIARRHEAL MORTALITY OF CHILDREN IN ASSOCIATION WITH HYDROGRAPHY IN BRAZIL

S. Leyk [a, b, *], T. P. Phillips [c, b], J. M. Smith [b], J. R. Nuckols [a, d]

[a] Fogarty International Center for Advanced Study in Health Sciences, U.S. National Institutes of Health, Bethesda, MD, USA
[b] Dept. of Geography, University of Colorado, Boulder, CO, USA - (stefan.leyk, jmsmith)@colorado.edu
[c] Colorado Center for Astrodynamics Research, Dept. of Aerospace Engineering, University of Colorado, Boulder, CO, USA - thomas.phillips@colorado.edu
[d] Dept. of Environmental & Radiological Health Sciences, Colorado State University, Fort Collins, CO, USA - john.nuckols@colostate.edu

KEY WORDS: Scale, Modelling, Spatial Infrastructures, Temporal, Spatial Epidemiology

ABSTRACT:

Is the pattern of childhood mortality due to diarrhea related to hydrography? To answer this kind of question we describe a method to examine associations between peak time of mortality from diarrhea in children less than 5 years for the period 1979-89 and relative location along the hydrologic regime of 8 major hydrographic regions in Brazil. In order to better understand the underlying heterogeneity found in mortality peak timing patterns a multiscale approach is used. Continuous geostatistical surfaces of mortality peak timing are created and validated based on these different spatial scales. We extract "mortality peak timing profiles" along the main streams within Brazil and test for possible trends in this peak timing variable in stream flow direction. Some first results demonstrate that there is a dominant trend of increasing mortality peak timing values downstream for most of the hydrographic regions. There are considerable differences between the spatial scales due to the higher degree of heterogeneity in peak timing at finer spatial scales, which remains hidden at coarser spatial scales. We found some deviations in the explored trends, which are possibly caused by a lack of underlying data and thus higher uncertainty in the underlying geostatistical model. Two hydrographic regions even show a trend into the opposite direction, upstream. Our approach allows for the formulation of interesting hypotheses regarding the capture of regional dynamics of diarrheal disease within various hydrographic regions. However, further research and refinement of assumptions is needed to improve this approach. One distinct advantage of the proposed approach is that the results are robust against effects of regional variation and non-linearity of relationships.

1. INTRODUCTION

The importance of an improved understanding of how and why water-borne diseases such as diarrhea spread has been widely recognized and has particular relevance for the Spatial Epidemiology community. There is strong demand to investigate the driving factors of spread and to establish a more effective surveillance system in order to explore successful preventive measures (Rushton, 2003; Pande et al., 2008; Chaikaew et al., 2009). This is of special priority in less developed places where the disease burden from enteric diseases is still very high (Fewtrell et al., 2005) but data to analyze disease spread and potential risk factors are scarce.

Recent efforts analyzed spatio-temporal patterns of diarrhea in order to explore regional dynamics of the disease and the main driving factors such as climatic variables and socio-economic conditions (Kelly-Hope et al., 2008; Jepsen et al., 2004). Frequently, such efforts rely on data from highly aggregated administrative reporting units (e.g., the scale of states or even larger administrative units). This results in a lack of statistical power of the analysis and a higher risk of ecological fallacy (Wakefield and Shaddick, 2005), which recognizes scale and thus aggregation as an important factor in understanding the heterogeneity of patterns and non-linear mechanistic relationships (Peters et al., 2004; Beale et al., 2008). Due to effects of regional variation in such relationships the use of global statistics and models is often impeded. Furthermore the arbitrariness of administrative analysis units in comparison to natural boundaries such as watersheds has raised discussions regarding the representativeness of the underlying statistical surface (Curriero et al., 2001).

The knowledge of fine-scale dynamics of diarrheal disease and driving factors is very limited. Hence advanced methods for the analysis of patterns of timing of diarrheal disease (spread) should allow for the investigation at different spatial scales in a geospatial framework and include natural boundaries. They should also ensure that results are robust against effects of non-linearity and regional variation of associations between the timing of a disease and environmental factors.

In this paper we present an approach to investigate the timing of annual maximum death cases (mortality peak timing) in children younger than five years in Brazil during the time period 1979-89

* Corresponding author. Address all inquiry to Stefan Leyk - stefan.leyk@colorado.edu

and its spatial association with relative location within hydrologic networks on a regional level. During the study period diarrheal mortality in Brazil showed strong seasonality and total mortality in children was relatively high since socioeconomic conditions were still under-developed. Our approach examines patterns of mortality peak timing at different spatial scales based on administrative reporting units and quantifies potential trends in the mortality peak timing variable along hydrological networks within major hydrographic regions of Brazil.
Our underlying hypothesis is that patterns of peak mortality with pediatric diarrhea as the causal factor are dependent on location of the geographic reporting unit in the hydrologic regime.

2. DATA

We obtained Brazilian mortality data from 1979-89 from the Ministério da Saúde do Brasil (http://www2.datasus.gov.br/DATASUS/index.php?area=02) classified as intestinal infectious disease (ICD-9 codes) and selected all deaths of children younger than 5 years of age. The death cases are recorded by municipality and with a temporal resolution of one month.

Spatial layers of administrative boundaries for sub-national levels as well as municipal localities and population densities were obtained from the Instituto Brasileiro de Geografia e Estatística (IBGE; http://www.ibge.gov.br/). We also used hydrographic data from the Agência Nacional de Águas (ANA; http://www2.ana.gov.br/). In order to carry out hydrological modeling we obtained 1 km spatial resolution elevation data from the Shuttle Radar Topography Mission (SRTM; http://www2.jpl.nasa.gov/srtm/).

3. METHODS

The approach described below consists of several steps and was based on some simplified assumptions. First, we extracted the peak timing values from the raw mortality data for each spatial scale (state, meso region and micro region, which are three nested levels of administrative division; micro regions consist of several municipalities) and calculated the mean peak values for the period 1979-89 based a continuous time scale. Next, exploratory spatial statistics were calculated to test for global spatial autocorrelation and spatial clustering in the mortality peak time variable. We then created and validated continuous geostatistical surfaces of mortality peak timing for all spatial scales. Finally we extracted the peak time values for stream locations in order to create "mortality peak timing profiles" along hydrologic main streams. These profiles were input to a final analysis of trends of mortality peak timing along the projected main flow direction within major hydrographic regions.

3.1 Extracting mortality peak timing

For each spatial scale (Figure 1 and 2a), we extracted the annual mortality peak months for the years 1979-89. A peak month is the month of a considered year in which the maximum of death cases (or mortality rate) was recorded. In this first experiment a peak month is valid if the maximum value is unique and greater than five deaths based on exploratory inspection. We calculated the mean peak timing only for spatial units that have more than four valid annual peak months during the considered time period 1979-89 and calculated inter-quartile ranges as variation measures. In order to create a continuous time scale that ignores the cut-off due to the end of the year we converted the peak month variable such that 01 September represents the starting point (1.0) and increases continuously (e.g., mid February has the value 6.5). Due to the above constraints but also caused by changes in administrative boundaries (e.g., increase of micro-regions between 1979 and 1989) and truly missing data, a certain proportion of spatial units (e.g., 126 of 557 micro-regions; 5 of 137 meso regions) remained without a mean mortality peak time value.

3.2 Exploratory spatial statistics

In order to test for global spatial autocorrelation in mortality peak timing at the different spatial scales we calculated global Moran's I (Moran, 1950). In addition we derived local indicators of spatial association (LISA), i.e., local Moran's I (Anselin, 1995) to test for the presence and type of significant spatial clusters and outliers in the mortality peak timing variable. The types of significant clusters (clustered low or clustered high values) can be used to identify regional differences in the spatial distribution of the data, which might be caused by various underlying processes. Because of the proportion of missing data at the scale of micro regions, the spatial structure of the polygon features was incomplete. We resolved this issue by assigning spatial weights based on polygon centroids and inverse distances to neighbours.

3.3 Geostatistical surfaces of mortality peak timing

In order to estimate mortality peak timing at each location – including the missing spatial units - we computed continuous geostatistical surfaces based on the polygon centroids for each spatial scale using Ordinary Kriging. The surfaces were created with a spatial resolution of 100km. First, model parameters (i.e., nugget, major range, partial sill, and lag size) were derived from a full model fit using all valid peak time values. Next the same parameterization was fit to a random selection of 65 percent of the centroid locations. In order to keep the model simple we avoided over-fitting and did not remove trends.

The geostatistical model was validated using the remaining 35 percent of the centroids. We calculated Spearman correlation coefficient and Kendall's Tau between the predicted (modeled) raster cells and the original mortality peak time values at the centroid locations. In addition we computed the Mean Absolute Error (MAE) to test for systematic bias. We iterated through this validation procedure 100 times.

Finally we selected out of the 100 processed geostatistical surfaces the ones for which the correlation coefficient was highly significant ($p < 0.01$) and calculated the mean peak time raster for each spatial scale. These mean values were used as approximations of mortality peak time at each location. As a final validation step we calculated the correlation between peak timing at all known centroids and the underlying mean peak time raster values.

3.4 Analyzing trends of mortality peak time along main streams

We tested for trends in the mean peak time rasters in (or against) flow direction of the hydrological network within each major hydrographic region and explored differences across underlying spatial scales.
We determined the location of the hydrologic network based on a flow accumulation greater than 10,000 pixels. A segment was defined as any individual stream section within the dendritic stream network occurring between an upstream starting point and a confluence, between two confluences, or between a confluence and the mouth of the river (or country border). Hence, segments can represent tributaries or sections of the main or feeder channels. We extracted the mean peak time raster values at the locations of each segment and created mortality peak time profiles. We calculated the difference between the peak time values at the first and last position of each mortality peak time profile (lowest and highest values of flow accumulation, respectively, along the segment),

These differences were recorded for each profile and weighted by the length of the profile to derive the average change in mortality peak time per km along the considered segment. This calculation was done for all profiles and aggregated for each major hydrographic region of Brazil. We also assessed the proportion of profiles with trends of increasing mortality peak timing downstream vs. upstream.

4. RESULTS

4.1 Exploratory spatial statistics

Global Moran's I statistics vary with changing spatial scale (Table 1). In general the finest resolution (micro-region) shows the lowest I value but the highest Z score in comparison to meso-regions and states. This could be an indication for local spatial autocorrelation of different directions (positive/negative) over the total area. In some of the individual years the I values were even close to zero, which could result from a balance of positive and negative local correlations.

The maps, which resulted from the spatial cluster analysis (LISA), for the three different spatial scales can be seen in Figure 1. At the coarsest resolution (states) there are significant spatial clusters of low values (LL) of mortality peak timing in the central western region and clusters of high values (HH) of mortality peak timing in the north-east region. These clusters are also identified but appear slightly expanded at the spatial scale of meso regions. Interestingly, significant outliers (a high value surrounded by low values (HL), or a low value surrounded primarily by high values (LH)) appear at the meso scale and connect the two aforementioned clusters. This continuum of significant cluster cores and outliers indicates a trend in local autocorrelation patterns (from LL to HH) as you move west to east from the interior to the coast. Finally the micro regions show the most differentiated and heterogeneous pattern. It can be seen that some clusters derived at the two coarser scales are relying on very few units at the micro region level. The transition between different types of clusters and outliers can also be identified at the scale of micro regions but the pattern appears much more differentiated and perforated. One point of attention is the most northern spatial unit at the meso region level that represents an HL outlier (Figure 1); it illustrates a high value of mortality peak timing surrounded by neighbors of low values.

	Spatial Scale		
	State	**Meso**	**Micro**
Nr. of years with significant I values	3/11	9/11	10/11
I (Z) for Median 1979-89	0.41 (3.20)	0.46 (10.96)	0.22 (22.22)
I (Z) for Mode 1979-89	0.35 (2.75)	0.32 (7.67)	0.12 (12.03)
I (Z) for Mean 1979-89	0.46 (3.52)	0.46 (11.00)	0.23 (23.52)

Table 1. Global Moran's I for different spatial scales (all significant; $p < 0.01$; I = I statistic, Z = Z-score)

Figure 1. Maps of spatial clusters of mortality peak timing from the LISA analysis for three spatial scales; states, meso and micro regions. Centers of significant spatial clustering and outliers are indicated as follows: HH = high-high, HL = high-low, LH = low-high, LL = low-low, NC = not clustered ($p < 0.05$). Data from uncolored administrative units did not meet the requirements for this analysis (see Methods 3.1) and were not included.

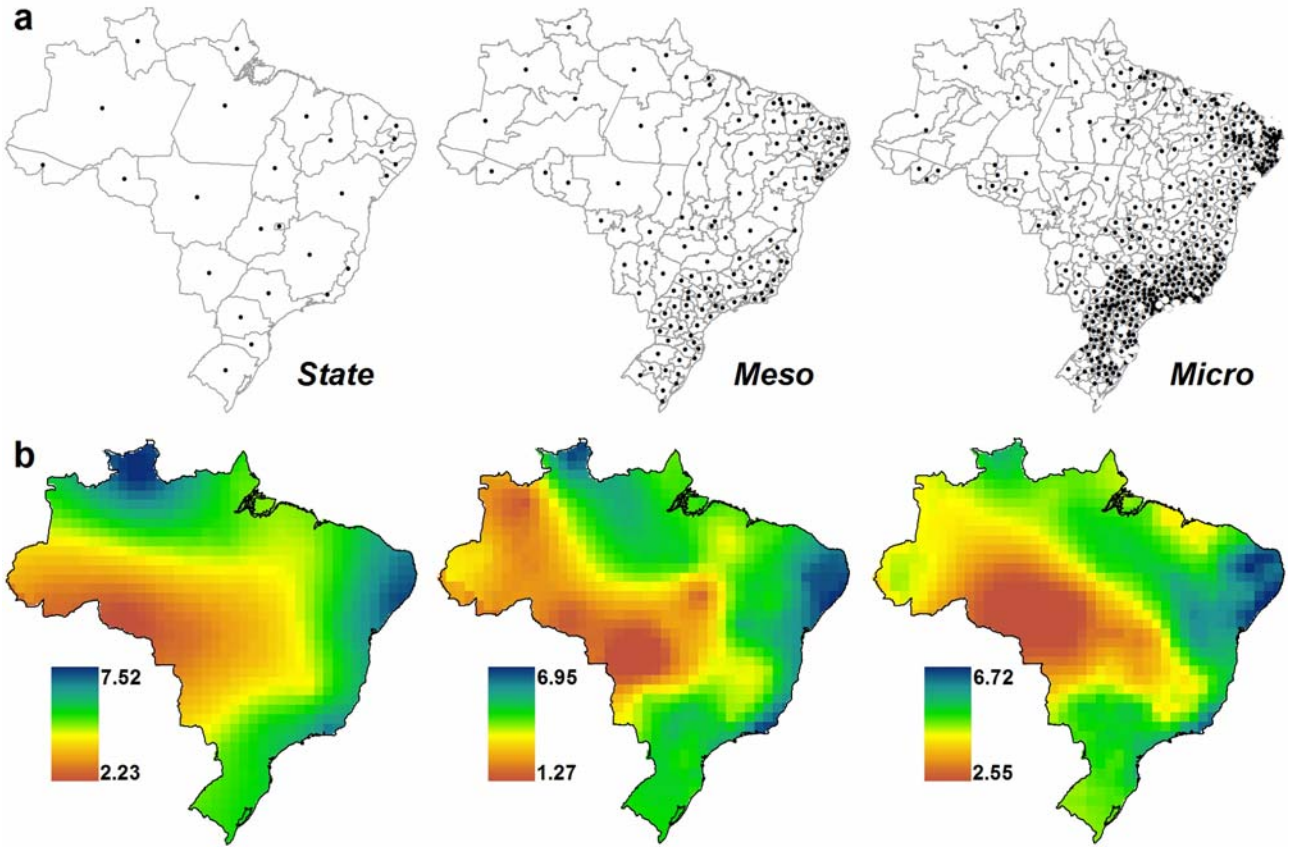

Figure 2. (a) Administrative units at three spatial scales. Centroids were used as input to the Kriging model. Polygons without centroids represent administrative units wherein mortality data did not meet the requirements for this analysis. Consequently these administrative units did not contribute to the kriging model. (b) Validated geostatistical surfaces of mortality peak timing based on Ordinary Kriging. The overall value range of 1.27 to 7.52 reflects the continuum of peak timing where September 1st is expressed as 1.0, October 1st is expressed as 2.0, etc.

4.2 Validation of geostatistical surfaces of mortality peak timing

The number of geostatistical surfaces for which the correlation tests between the modeled raster values and original values at left out centroid locations varied between the different scales of analysis. Out of the 100 model runs 43% at the state level, 94% at the meso level and 98% at the micro level were tested highly significant ($p < 0.01$) and were thus used for calculating the mean peak time raster. The final validation step of the mean peak time rasters resulted in correlation (Spearman) and Mean Absolute Error of 0.96 (MAE = 0.244) for state, 0.90 (MAE = 0.445) for meso and 0.72 (MAE = 0.693) for micro regions. The three validated continuous surfaces from this procedure are shown in Figure 2b. In addition variance surfaces were calculated in order to quantify the expected accuracy of the models in different regions (not shown). They indicate highest uncertainty where reporting units do not have valid peak time values.

4.3 Trends of mortality peak time along main streams

A visual interpretation of the mortality peak timing surface overlayed by the hydrologic network provides some first indication of a trend of increasing (delayed) peak time downstream within the hydrological network of Brazil (Figure 3a). A more detailed inspection of these high-resolution profiles at the scale of micro regions (Figure 3b) shows that there is a general gradient of increasing mortality peak timing from the upper elevation of the hydrographic regions towards the most downstream segment (confluence of the major river with the ocean or country border), especially for the major rivers in Brazil (hydrographic regions 1, 2, 4, 6, 8 in Figure 3b). This gradient is less obvious for the coarser spatial scales.

The magnitudes and directions of trends as derived from first calculations vary across the different spatial scales for some hydrographic regions (e.g., region 2 and 3). For example in hydrographic region 2 an average change per km in mortality peak timing along segments of 0.0187 (downstream) for state, 0.0048 (upstream) for meso and 0.0491 (downstream) for micro regions could be found. Other hydrographic regions showed similar values and constant directions of average changes in mortality peak time across scales (for example 1, 4 and 5 showed all downstream trends of increasing mortality peak timing). These differences are also reflected by varying proportions of profiles with a trend of increasing mortality peak time downstream across spatial scales.

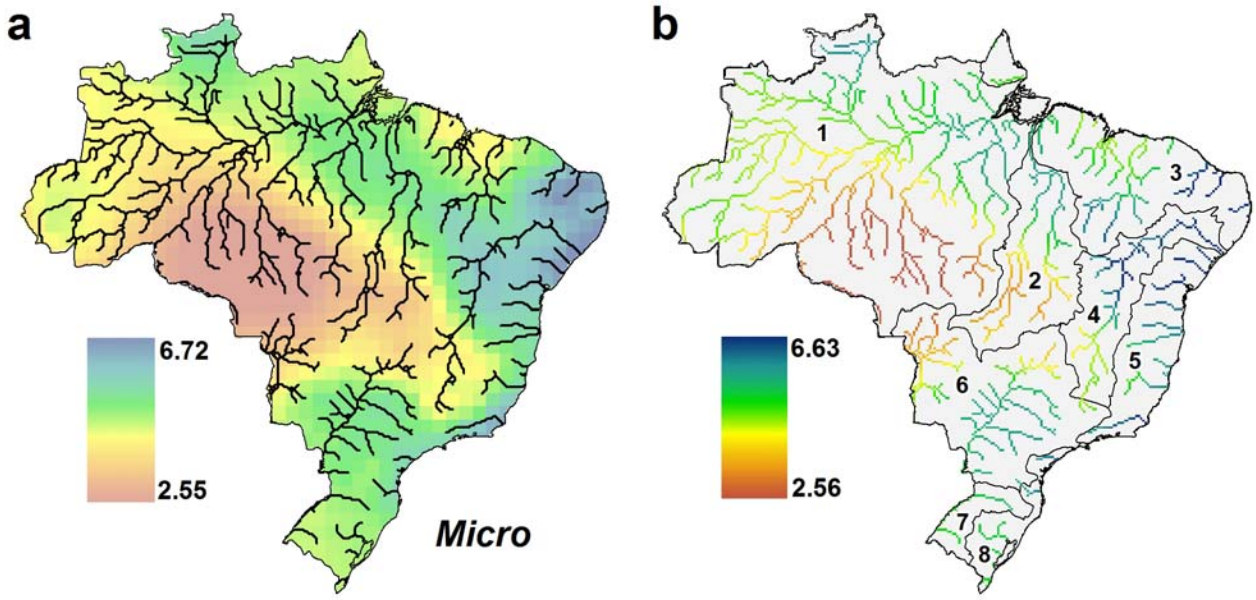

Figure 3. (a) Hydrologic network overlaid at the geostatistical surface of mortality peak timing derived for the micro region level. (b) Extracted profiles of mortality peak time values along the main streams within major hydrographic regions of Brazil. The black boundary line indicates named hydrographic regions as follows: 1 = Amazônica; 2 = Tocantins-Araguaia; 3 = Atlântico Nordeste Ocidental, Atlântico Nordeste Oriental, Parnaíba; 4 = São Francisco; 5 = Atlântico Sudeste, Atlântico Leste; 6 = Paraná, Paraguai; 7 = Uruguai; 8 = Atlântico Sul

In regions where reporting units are missing due to the described constraints of this analysis it can be seen that trends are less obvious or switch from downstream to upstream. For example in the western part of hydrographic region 1 a considerable proportion of increasing mortality peak time values upstream was identified at the micro scale despite a general dominant trend downstream. This could be due to model misspecifications caused by missing reporting units but

also by unstable mortality rates in these sparsely populated regions. These effects seem to be smoothed at the coarsest, more aggregated, spatial scale (state).

5. DISCUSSION AND CONCLUDING REMARKS

We described a first attempt to spatially explore potential associations between peak timing of mortality from pediatric diarrhea in Brazil (1979-89) and the relative location along the main streams of the hydrologic network within major hydrographic regions. Based on a set of simplified assumptions and constraints we found dominant trends of increasing mortality peak timing from the watershed divide to the most downstream segment along the main streams.

In some hydrographic regions we observed changes in trend direction when comparing the different spatial scales. This effect seemed to occur in regions where reporting units were missing due to the constraints of this analysis or where the hydrologic network mainly consisted of short rivers. The latter finding could be an indication that a trend is difficult to identify for too short segments due to the coarse resolution of the mortality peak timing variable of 100km. The geostatistical models based on the finest scale (micro region) showed the highest degree of heterogeneity in mortality peak timing and strongest trends downstream but also considerable areas of higher variance, which could be caused by unstable and missing mortality peak timing values for these fine scale reporting units. The assumptions and constraints used here will be revisited to test this approach for sensitivity and to allow for more complexity involved.

The trend analysis described here focused on zones along the main streams according to our hypothesis that diarrheal mortality peak timing is related to hydrography. The use of geostatistical models based on centroid locations could be considered a weakness since a number of administrative units lacked adequate data. Nevertheless, creating continuous raster surfaces based on validated geostatistical models is the most reliable procedure one could carry out in such a situation.

We did not account for under-reportings, which are evident for Brazilian mortality data during the considered time period (Castro et al., 2010), and this represents a potential deficiency of our analysis. However to simplify the analysis we assumed that peak timing as a variable should be relatively robust if a minimum number of cases can be found even in the presence of under-reportings. This issue will be further investigated in the near future.

Future research will also focus on the incorporation of more advanced higher resolution dasymetric refinement to concentrate the analysis on sub-regions most important for observing diarrheal disease dynamics and to include anthropogenic processes relevant for exposure to pathogens associated with enteric disease (e.g. population centers). This will also include the improvement of the geostatistical model by taking into account additional variables such as elevation and climatic control variables as well as the evaluation of other geostatistical techniques in order to test for replicability of the findings.

REFERENCES

Anselin, L., 1995. Local Indicators of Spatial Association – LISA. *Geographical Analysis*, 27, pp. 93-115.

Beale, L., Abellan, J. J., Hodgson, S., Jarup, L., 2008. Methodologic Issues and Approaches to Spatial Epidemiology. *Environmental Health Perspectives*, 116, pp. 1105–1110.

Chaikaew, N., Tripathi, N. K., Souris, M., 2009. Exploring spatial patterns and hotspots of diarrhea in Chiang Mai, Thailand. *International Journal of Health Geographics*, 8, pp. 36.

Curriero, F. C., Patz, J. A., Rose, J. B., Lele, S., 2001. The Association Between Extreme Precipitation and Waterborne Disease Outbreaks in the United States, 1948–1994. *American Journal of Public Health*, 91, pp. 1194-1199.

Fewtrell, L., Kaufmann, R. B., Kay, D., Enanoria, W., Haller, L., Colford, J. M., 2005. Water, sanitation and hygiene interventions to reduce diarrhea in less developed countries: a systematic view and metanalysis. *Lancet Journal of Infectious Diseases*, 5, pp. 42-52.

Jepsen M. R., Simonsen J., Ethelberg J. S., 2004. Spatio-temporal cluster analysis of the incidence of Campylobacter cases and patients with general diarrhea in a Danish county, 1995-2004. *International Journal of Health Geographics*, 8, pp. 11.

Kelly-Hope, L. A., Alonso, W. A., Thiem, V. D., Canh, D. G., Anh, D. D., Lee, H., Miller, M. A., 2008. Temporal Trends and Climatic Factors Associated with Bacterial Enteric Diseases in Vietnam, 1991–2001. *Environmental Health Perspect*ives 116, pp. 7–12.

Moran, P. 1950. Notes on Continuous Stochastic Phenomena. *Biometrika*, 37, pp. 17–33.

Pande, S., Keyzer, M. A., Arouna, A., Sonneveld, B. G. 2008. Addressing diarrhea prevalence in the West African Middle Belt: social and geographic dimensions in a case study for Benin. *International Journal of Health Geographics*, 7, pp. 17.

Peters, D. P., Pielke, R. A., Bestelmeyer, B.T., Allen, C. D., Munson-McGee, S., Havstad, K. M. 2004. Cross-scale interactions, nonlinearities, and forecasting catastrophic events. *Proceedings of the National Academy of Sciences of the United States of America*, 101, pp. 15130-15135.

Rushton, G. 2003. Public Health, GIS, and Spatial Analytic Tools. *Annual Review of Public Health*, 24, pp. 43-56.

Wakefield, J., Shaddick, G. 2005. Health-Exposure Modelling and the Ecological Fallacy. *Biostatistics*, 1, pp. 1–19.

Castro M. C. and Simões C. C. 2010. Spatio-Temporal Trends of Infant Mortality in Brazil. *Proceedings of the Annual Meeting. Population Association of America* (PAA) 2010.

ACKNOWLEDGEMENTS

Funded in part by the Bill and Melinda Gates Foundation and the Fogarty International Center as well as by an Interagency Personnel Agreement between Division of International Epidemiology and Population Studies, Fogarty International Institute, National Institutes of Health and University of Colorado at Boulder.