Published in final edited form as:

Spat Spatiotemporal Epidemiol. 2012 June ; 3(2): 95–105. doi:10.1016/j.sste.2012.04.003.

Robust Assessment of Spatial Non-Stationarity in Model Associations Related to Pediatric Mortality due to Diarrheal Disease in Brazil

Stefan Leyk^{a,b,*}, Petra U. Norlund^a, and John R. Nuckols^c

^aDepartment of Geography, University of Colorado, Boulder, CO, USA

^bDivision of International Epidemiology and Population Studies, Fogarty International Center, National Institutes of Health, DHHS, Bethesda, MD, USA

^cDepartment of Environmental and Radiological Health Sciences, Colorado State University, Fort Collins, CO, USA

Abstract

Approximately 1.5 million people, mostly children, die annually due to disease attributed to diarrhea reflecting urgent needs for improved understanding of associations between the disease and potential risk factors. Numerous epidemiological studies found spatially varying (non-stationary) disease associations attributable to changing geographic or demographic context. Spatial non-stationarity implies that average relationships from statistical models fitted to the whole study area might be inappropriate since they do not reflect local conditions. Spatial modeling techniques such as Geographically Weighted Regression (GWR) have limitations in providing statistically robust analysis of spatial non-stationarity. Thus, there is a need for development or expansion of modeling techniques to address this issue.

Using data for pediatric diarrheal mortality in Brazil in 2000, and different risk factors, we develop an analytical framework to determine regions of similar (stationary) local associations by combining GWR and max-p regionalization. We fit statistical models to these regions, and compare goodness-of-fit and regionally varying coefficients to the national-scale model measures. The proposed framework allows us to examine (a) impact of non-stationarity for regions of different geographic extent with acceptable statistical power, (b) the explanatory power of each risk factor in each region, and (c) if these regions reflect changing data quality or truly existing variations in putative associations.

Keywords

Pediatric diarrhea; spatial non-stationarity; geographically weighted regression; regionalization; disease modeling

*Corresponding author: Department of Geography, University of Colorado, UCB260, Boulder, CO, 80309, USA; stefan.leyk@colorado.edu.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

1 Introduction

Research on modeling diarrheal disease traditionally explored statistical associations with potential environmental, climatic and socio-economic risk factors using non-spatial statistical models (Strina et al., 2003; Kelly-Hope et al., 2007) which assume geographically invariant processes and thus stationary associations. Several recent studies reported on the existence of spatially autocorrelated and clustered outcome data related to diarrheal disease (Pande et al., 2008; Jepsen et al., 2009; Chaikaew et al., 2009; Alonso et al., 2012), which has implications for the analytical procedure (Elliott and Wartenberg, 2004; Beale et al., 2008). In addition to such observed geographical pattern spatial variation, ‘spatial non-stationarity’ (SNS) of statistical associations (Brunsdon et al., 1996) between disease outcome and risk factors can occur. The occurrence of spatial non-stationarity implies that global statistical models over a given study area can become inappropriate since they represent average relationships that are not transferrable to subregional or local conditions (Cheng et al., 2011). For example, associations between disease and potential risk factors can vary across subregions of a study area due to variation in societal contexts (Nakaya et al., 2005), environmental or climatic conditions. SNS has been rarely investigated in the context of modeling diarrheal disease (e.g., Carrel et al., 2011). However, recent studies on SNS have been conducted concerning other disease outcomes. These include studies of tick-borne diseases (Wimberly et al., 2008; Lin and Wen, 2011), HIV/AIDS (Nakaya et al., 2005), childhood obesity (Edwards et al., 2010), and breast cancer (Mandal et al., 2009; Tian et al., 2011). These studies applied geographically weighted regression (GWR) models which allow estimation of local coefficients for each observation using spatially weighted observations located in the local neighborhood (‘kernel’), the size of which is defined by a bandwidth measure (Brunsdon et al., 1996; Fotheringham et al., 2002). In most cases high degrees of variation in local associations between disease and risk factors were found, further indicating the use of global statistical models in analyzing the data could produce erroneous results.

While the above body of research witnesses the importance of spatial non-stationarity in disease-related associations, available analytical approaches such as GWR are subject to fundamental critiques concerning a lack in robustness for statistical inference (O’Sullivan and Unwin, 2010). GWR models can suffer from local over-fitting and can show patterns of induced spatial non-stationarity as a result of spatial weighting of the observations in each local regression (Cho et al., 2010). Further problems that have been identified are the instability of coefficient estimates as a function of bandwidth (Farber and Páez, 2007) and the occurrence of multicollinearity of the local coefficient estimates (Wheeler and Tiefelsdorf, 2005; Griffith, 2008).

The above problems, combined with the complexity of GWR and the subjectivity introduced by parameter settings are serious impediments for objective evaluation and for communicating the results to the health community and policy makers. In order to overcome such impediments the analysis of spatial non-stationarity needs to be incorporated into a statistical framework that integrates principles of geographically weighted regression techniques in a reliable and objective way. Such a framework would help foster the

understanding of effects and implications of spatial non-stationarity in disease model associations at broader geographic scales. This article describes an analytical framework designed to incorporate GWR models (computed for different bandwidths) for guiding a regionalization process that determines the underlying subpopulations (i.e., regions within the broader study area) to be used to fit regional-scale models. This procedure allows the user to minimize adverse effects of spatial (or regional) non-stationarity when using one global model for the whole study area while ensuring acceptable statistical power for regional models. We evaluate the approach by modeling relationships between pediatric diarrheal mortality recorded in Brazil in 2000 and the known or presumed risk factors, such as population density, socio-economic status, percent illiterate population as a surrogate for educational status and household access to water.

2 Data

Population and socioeconomic data

The geographic extent (i.e., 'national scale') of our study area is the geopolitical boundary of Brazil. From the Instituto Brasileiro de Geografia e Estatística (IBGE) we obtained spatial layers of polygons representing administrative boundaries of the 558 Census Micro Regions (CMRs) used to report census information for the year 2000 in Brazil (<http://www.ibge.gov.br/>). For each CMR we obtained population statistics by age from the same year as well as estimates of urban versus rural population, gross domestic product and per-capita income. We calculated population density of CMRs based on census population estimates and area of each CMR.

From the Ministério da Saúde do Brasil, we obtained year 2000 data on the number of households in each CMR with access to water, access to sanitation and plumbing facilities, as well as on illiteracy as a proportion of the population older than 15 years that is illiterate (all for the year 2000).

Mortality outcome data

From the Ministério da Saúde do Brasil, we obtained year 2000 Brazilian mortality data for children younger than five years of age for which the primary cause of death was classified as related to intestinal infectious disease (ICD-9 codes) (<http://www2.datasus.gov.br/DATASUS/index.php?area=02>). In the dataset, monthly mortality is reported by Municipality, a census reporting unit of the Brazilian government with an approximate median area of 420 km² (a total of 5,507 Municipalities in 2000 (IBGE)). We aggregated the number of deaths to the same reporting unit as for our population, socio-economic and environmental risk factor data - the Census Micro Regions (CMR) of Brazil (Median area 5480 km²; Figure 1) each of which contains multiple Municipalities. Using age-classified population statistics from IBGE, we were able to calculate mortality rates for the year 2000 such as deaths attributed to intestinal infectious disease per 100,000 children or deaths per capita of children under the age of five, for each CMR (Figure 1).

3 Methods

The method described here consists of four subsequent steps:

1. Computing non-spatial global statistical ‘baseline’ models for CMR-level mortality attributed to intestinal infectious disease in children under age five, standardized by population at risk for the whole of Brazil (national scale), to identify statistically significant risk factors; testing model residuals for spatial autocorrelation and clustering;
2. Computing generalized GWR models for Poisson distributed data and mapping local coefficients for a broad range of bandwidth measures;
3. Defining contiguous homogeneous regions (sub-national or regional scale) of varying size using the max-p region algorithm; taking local GWR coefficients as input from two bandwidth scenarios (small and large bandwidth) to assess effects of multicollinearity in GWR coefficients on the regionalization output; and
4. Fitting models for each of the regions delineated by the max-p algorithm, relating mortality due to intestinal infectious disease to the risk factors identified in Step (1); Finally identifying the variation in regional model coefficients as well as comparing regional and national-scale model coefficients and goodness-of-fit measures.

3.1 National-scale statistical ‘baseline’ model

In order to define a reference point for evaluating model performance we fit a ‘baseline’ generalized linear model (GLM) adjusted for Poisson distributed count data of diarrhea related deaths for the whole of Brazil (national scale) and standardized by the population at risk, CMR level population of children of less than five years of age. We recorded model coefficients (β_{na}) for each variable (risk factor), tested for statistical significance, and calculated the Akaike Information Criterion (AIC) (Akaike, 1974). We tested model residuals for over-dispersion and spatial autocorrelation, which would indicate a non-random error structure. We also tested for significant local clusters in model residuals based on local indicators of spatial associations (LISA) (Anselin, 1995) using Rook contiguity (i.e., CMRs are adjacent if they share a boundary) for creating the spatial weights matrix (999 permutations; $p < 0.05$). Clusters of high and low residual values at the CMR level may indicate spatial variation in model performance. We selected independent variables that were significant ($p < 0.05$) and tested them for multicollinearity (Belsley et al., 1980). Of the significant variables that were highly correlated with each other we selected the most significant ones which were percent illiteracy (*illit*) and the number of households with access to drinking water (*wtr_h*) for use in our study; both variables are considered known risk factors for diarrheal disease. We held these as the constant predictor variables in the evaluation of regional versus national geographic scale models in our study.

In order to examine the co-variation of the (standardized) value of mortality (here: mortality rate per 100,000 children of less than five years) at a given CMR with the average of neighboring values of each of the selected risk factors we computed the bivariate LISA. This is an extension of the univariate form of LISA to investigate the spatial association between two variables (999 permutations; $p < 0.05$).

3.2 Generalized GWR and local coefficient mapping

GWR models adapt the idea of varying coefficient models (Cleveland et al., 1991; Hastie and Tibshirani, 1993) to the spatial case and therefore allow for local, spatially weighted estimations of coefficients (i.e., weighted for each observation) to account for spatial variation of these coefficients across a given study area. GWR has been commonly used as an exploratory mapping tool (Wheeler and Páez, 2010).

Using the two most statistically significant predictor variables for CMR level mortality attributed to diarrheal disease in the year 2000, percent illiteracy (*illit*) and household access to water (*wtr_h*) (Section 3.1), we applied a form of Generalized GWR (GGWR) (Fotheringham et al. 2002) that allows us to adjust for Poisson distributed outcome (count) data standardized by the population at risk. This method has been applied in studies on homicide mortality in Brazil (Araújo et al., 2010) and cervical cancer in England (Cheng et al., 2011) to demonstrate its suitability for modeling with non-normal data distributions. The statistical foundation of Poisson adjusted GGWR (or 'GWPR') models can be found in Cheng et al. (2011) and Nakaya et al. (2005).

The calculation of the bandwidth, which defines the area (or kernel) for local estimation in GWR, represents an ongoing research question (Cho et al. 2010). While global bandwidth measures result in the same kernel size at each location, adaptive bandwidths are determined by the proportion of neighboring observations to be included and local kernels may therefore vary in size. We attempted to derive the bandwidth parameter using the most common approaches, i.e. cross-validation or using the Akaike Information Criterion (AIC) (Akaike, 1974, Brunsdon et al. 1996). However, these approaches did not provide meaningful results in our study likely due to high degrees of variation in data density (Figure 1). For this reason we employed a range of adaptive bandwidth measures (proportions of CMRs between $0.025 = 14$ CMRs and $0.25 = 139$ CMRs) to examine the sensitivity of the resulting spatial distributions of local GWR coefficients to changes in the bandwidth measure. In particular we intended to test the effect of increasing the number of CMRs used for local estimation on the stability and multicollinearity of regression coefficients (Farber and Páez, 2007) - starting with very few (14) CMRs and progressing stepwise to many (139) CMRs. In order to calculate the spatial weights for neighboring CMRs within the kernels we used the Gaussian function as the 'kernel function'. For each bandwidth we calculated and mapped summary statistics for the model coefficients (median, upper and lower quartiles, and minimum and maximum values) in order to quantitatively and visually assess the variation (spatial nonstationarity) in local model associations between mortality and both selected predictor variables, *illit* and *wtr_h* (Section 3.1). Based on these results, we selected two bandwidth scenarios for use in the subsequent regionalization (Section 3.3) and computation of models at the regional scale (Section 3.4). We selected one bandwidth that uses very few CMRs (i.e., 0.025 or 2.5%) and shows high degrees of variation (resulting in lower collinearity in coefficients), and one bandwidth that uses higher numbers of CMRs (i.e., 0.1 or 10%) showing lower degrees of variation, but higher collinearity in the GWR model coefficients.

3.3 Regionalization based on local model coefficients

For each of the two selected bandwidth scenarios we used the GWR coefficients to derive homogeneous regions of CMRs across our study area of the whole of Brazil that were spatially contiguous, exclusive (each area is assigned to only one region), and constrained (by a threshold value for a regional attribute). Since there is no theoretical basis to predetermine the number of regions to be generated, an approach was needed that identifies homogeneous regions endogenously dependent on regional threshold values (e.g., the minimum number of CMRs to be included in each region) and the spatial distribution of GWR coefficients. This endogeneity-feature cannot be provided by most existing regionalization approaches (e.g., Murtagh, 1985; Shirabe, 2005), which have seldom been tested for GWR coefficients (an exception being Wimberly et al., 2008, using k-means cluster analysis). We used the max-p region approach (Duque et al., 2007) provided in PySAL (Rey and Anselin, 2007) (<http://geodacenter.asu.edu/pysal>), a region swapping technique that is designed to maximize the number of endogenously determined regions of CMRs with similar area attribute values (i.e., GWR coefficients), thus minimizing heterogeneity.

Regions were computed using the CMRs as input areas, the matrix of estimated GWR coefficients for *wtr_h* and *illit* (Section 3.2) as attributes on each CMR, and the minimum number of CMRs in each region as regional threshold value ('floor constraint'). Different floor constraints were tested in order to examine their influence on the shape and location of the generated regions. These different floor constraints were derived from a simple power analysis that assumed a medium size effect ($d = 0.35$) as recommended by Cohen (1988, 1992) at a significance level (α) of 0.05. We determined sample sizes required for varying levels of statistical power (i.e., 80%, 95%, 97.5% and 99%), which resulted in 50, 80, 100 and 120 CMRs, respectively, and used these values as regional constraints. This procedure ensured that even the smallest regions still contained acceptable numbers of CMRs (50 CMRs or more) allowing for robust statistical models (Section 3.4).

Since max-p regions were created based on pairwise similarity tests (Duque et al., 2007) lower degrees of spatial variation (higher similarity) in local mortality-risk factor associations would be expected. We inspected the performance of max-p under varying floor constraints by observing whether resultant regions showed variation in size. If the majority of regions consist of the exact number of CMRs used as the floor constraint this could indicate that CMRs were possibly forced to join regions in order to satisfy the required size constraint resulting in a poor performance of max-p (i.e., regions that are less homogeneous). We found the regionalization using only 50 CMRs as floor constraint to be the only one in which regions varied in number of CMRs (Table 2) for both selected bandwidths (Section 3.2), therefore indicating a good performance of the max-p approach in capturing regions with internal homogeneity and external heterogeneity. Hence, we used this regionalization in the development of subsequent regional models for comparison to results from a national-scale model (Section 3.4).

3.4 Regional-scale models and comparison to national-scale models

The max-p regions that were generated with a floor constraint of 50 CMRs (Section 3.3) were used to recompute regional Poisson-adjusted GLMs (i.e., models were fit to the CMRs within these regions). We fitted models to max-p regions that were generated for the two selected GWR bandwidth scenarios, 2.5% and 10% of CMRs (Section 3.2). We performed goodness-of-fit tests for the regional-scale models using AIC (Akaike, 1974) and compared the results to a regionally adjusted measure of goodness-of-fit of the national-scale model (Section 3.1); a test that has recently been demonstrated in the context of migration (Leyk et al., in press) and was adjusted for this study. While AIC is commonly used to compare model fit for identical populations using different sets of independent variables, here the objective was to compare the goodness-of-fit of two models that use the same set of independent variables, but whose coefficients were estimated from fitting the models to regions of different geographic extent: the geographic boundary of an entire country ('national scale'; β_{nat}) versus subregions that are defined using the max-p approach ('regional scale'; β_{reg}) (Section 3.3). Given the maximum likelihood estimate L and the number of model parameters k AIC is calculated as (Akaike, 1974):

$$AIC = -2\log(L) - 2k \quad (1)$$

The log-likelihood function for the model with parameters β using n independent observations can be calculated as:

$$\log L(\beta) = \sum_{i=1}^n (y_i \log(\mu_i) - \mu_i) \quad (2)$$

where μ_i is the fitted response value for observation y_i . In order to compare the goodness-of-fit of the national-scale model with a regional-scale model, response values μ from the national-scale model (Section 3.1) were selected only for the subset of observations (CMRs) used for fitting the regional-scale model. This subset of the national-scale model's response values (μ) was used to calculate the maximum likelihood estimate L and, subsequently, AIC. For each region we then calculated the AIC for the regional-scale model and compared it to the AIC from the national-scale model applied to the same region and measured the difference in AIC. We calculated the percent change in AIC across all regions of each of the max-p regionalization outcomes. We compared coefficients from national-scale and regional-scale models (β_{nat} and β_{reg}) and calculated ranges of regional-scale model coefficients in relation to the national-scale coefficient value, a technique that has recently been applied for comparing model coefficients estimated at different temporal scales (Leyk et al., 2011).

In summary, the process outlined above systematically examines (i) whether GWR-driven regionalization has an effect on statistical inference and associations when modeling mortality attributed to intestinal infectious disease, (ii) whether a change in bandwidth used in GWR results in a different regionalization with possible effects on the subsequent

regional models, and (iii) spatial non-stationarity in statistical associations between mortality attributed to intestinal infectious disease and the two risk factors as estimated from regional-scale models with acceptable sample sizes and statistical power.

4 Results

4.1 Statistical baseline model on the national scale and diagnostics

The results from the national-scale Poisson GLM using illiteracy (*illit*) and household water access (*wtr_h*) as risk factors (β_{nat}) are shown in Table 2. Both variables and the model intercept tested highly significant with low standard errors of coefficient estimates. We found a positive relationship between mortality and illiteracy, and a negative relationship between mortality and household water access. We found highly significant global spatial autocorrelation in the model residuals ($I = 0.12$; $p < 0.001$) (Figure 2a). We also observed significant local clusters of low (LL) and high (HH) model residuals based on LISA - suggesting statistically significant clusters of over- and underestimations in southeastern/eastern and northeastern Brazil, respectively (Figure 2b).

The bivariate LISA maps in Figures 3a and b illustrate spatial associations between mortality and the two selected risk factors, *wtr_h* and *illit*. We found statistically significant clusters of high mortality and low values of household access to water, *wtr_h*, in neighboring CMRs (HL) in northern Brazil, and clusters of low mortality and high values of *wtr_h* in the neighboring CMRs (LH) in southeastern Brazil (Figure 3a). We observed a considerable number of bivariate clusters of low mortality and low values of *wtr_h* (LL) in neighboring CMRs. In Figure 3b clusters of high mortality and high illiteracy (HH) in the northern part, as well as clusters of low mortality and surrounding low illiteracy (LL) in the southern part, can be observed. We also observed statistically significant clusters of low mortality and high illiteracy (LH) in neighboring CMRs in the northeast and northwest. Similarly, there are significant clusters of high mortality and low illiteracy in neighboring CMRs in southern Brazil (HL).

4.2 Local estimation using GWR and mapping of GWR coefficients

The maps in Figures 4 and 5 show the resultant classifications of CMR-level local GWR coefficients of the two risk factors of interest (*illit* and *wtr_h*) estimated using different adaptive bandwidths (proportions between 0.025 to 0.25) (Section 3.2). The coefficient values, which correspond to the class breaks in Figures 4 and 5, can be found in Table 1. The coefficient surfaces for illiteracy (Figure 4) show fine-scale variations of abruptly changing local associations in the eastern/southeastern parts of the country, particularly for the smallest bandwidths (proportions of 0.025 and 0.05); we observed negative local coefficients (Table 1) indicating weak statistical relationships. With increasing bandwidths, surfaces appear increasingly smooth. The class of smallest estimated coefficients in the central/northwestern part of the country appears very consistent across all bandwidths tested.

The GWR coefficient surfaces for the variable *wtr_h*, household access to water (Figure 5), show different distributions (due to the negative global association) but follow similar trends

i.e., fine-scale variations at small bandwidths and smoother partitions with increasing bandwidths (for greater than 0.15) (Figure 5).

The interquartile range in coefficient values almost universally decreased with increasing bandwidth (Table 1) for both predictive variables.

4.3 Region-specific models and goodness-of-fit test results

The regional-scale modeling results for max-p regions generated using a floor constraint of 50 CMRs for both bandwidth scenarios, proportions of CMRs of 0.025 and 0.1, are shown in Table 2. Rows in Table 2 refer to region labels in the maps (A–H) in Figure 6.

While coefficients of both variables show considerable degrees of variation across regions (β_{reg} in Table 2), household access to water (wtr_h) tested significant in most regions for both bandwidth scenarios (not significant in regions D, which are almost identical in location by use of either of the two bandwidths). Illiteracy tested insignificant in several regions (A, B and H for bandwidth = 0.025 and A, F and H for bandwidth = 0.1), but was found to be highly significant in the other regions (Table 2), suggesting that associations between pediatric mortality attributed to intestinal infectious disease and illiteracy vary considerably across regions.

Based on the goodness-of-fit test (Section 3.4), we found improvements (i.e., decreases in AIC, see Table 2) for both adaptive bandwidth scenarios used in GWR in all regional-scale GLMs when compared to the national-scale GLM. The average decrease in AIC across all max-p regions (Section 3.3) was 10.27% for the bandwidth scenario 0.025 and 8.56% for the bandwidth scenario 0.1.

5 Discussion and Concluding Remarks

In this study, we present an approach to evaluating spatial non-stationarity in statistical associations between pediatric mortality attributed to intestinal infectious disease and known risk factors (household access to water supply (wtr_h) and prevalence of illiteracy in the population ($illit$). We demonstrate this analysis by comparing models at the national (or countrywide) versus regional scale in Brazil. Our results demonstrate substantial variations in goodness-of-fit measures (AIC_{reg}) as well as coefficient values (β_{reg}) as a function of bandwidth used in the GWR models. Relative to the coefficient for the national-scale model (β_{nat}), we observed regional coefficients ranging from $\beta_{nat}/17$ to β_{nat}^*7 for wtr_h and $-\beta_{nat}/4$ to β_{nat}^*12 for $illit$ when using bandwidth = 0.025. For bandwidth = 0.1 the ranges were $\beta_{nat}/2$ to β_{nat}^*7 for wtr_h and $-\beta_{nat}/1.5$ to β_{nat}^*5 for $illit$ (Table 2). The decrease in regional-scale model AICs as compared to the adjusted national-scale model AIC was between 2.3% and 20.6% for bandwidth = 0.025 and between 5.5% and 17.7% for bandwidth = 0.1. Overall, the model results indicate expected trends i.e., decreased mortality where there is increased access to water, and a positive relationship between mortality and illiteracy. However, the above ranges illustrate substantial variation in these coefficient values across the regions that decreases for increasing bandwidths used in the GWR model (Table 2). These variations and the occurrence of non-significant regional-scale coefficients, indicating that the considered risk factor is less important in these regions, would remain

hidden if only the national-scale model was used. These findings have important implications in terms of public health if such models are to be used to allocate resources and plan intervention and prevention strategies for disease control.

The above findings also demonstrate a need for a framework to determine regions of similar local disease-risk factor associations and to construct accurate and robust models for these regions. We present and evaluate such a framework in this paper (Figure 7).

Instead of estimating coefficients for each observation, which is typically done in GWR, we systematically determined the minimum size of regions needed to establish statistical associations with an acceptable statistical power (minimum 80%), established a priori, in order to avoid above described problems related to GWR. Results of our evaluation indicate that our proposed method overcomes shortcomings of commonly used local estimators, in particular the lack of robustness for statistical inference (O'Sullivan and Unwin, 2010). For example, we were able to assess spatial variation in coefficient values while at the same time these coefficients can be tested for significance and the goodness-of-fit can be calculated within the described GLM framework, hence the importance of single risk factors in different regions can be objectively evaluated.

Furthermore, by applying this methodology we were able to account for non-stationarity, robustness and collinearity in model coefficients observed by similar geographic-based studies of disease associations (Wheeler and Tiefelsdorf, 2005; Nakaya et al., 2005). For example, we used GWR as an early exploratory step to guide the max-p regionalization with the expectation that regions of high similarity in local associations (i.e., GWR coefficients) are identified. Within these regions it should then be possible to fit more robust statistical models with lower degrees of spatial non-stationarity in target associations. To this end, we examined how instability (Farber and Páez, 2007) and multicollinearity (Wheeler, 2007; Griffith, 2008) of GWR coefficients influence the regionalization of modeling units (i.e., location, shape and size) by testing a range of bandwidth measures (Table 1; Figures 4 and 5). We found that a small bandwidth (proportion 0.025), which in GWR implies very few local observations, showed less compact regions (Figure 6a) and higher degrees of variation in regional-scale model coefficients compared to a greater bandwidth value (Table 2). We also found that the max-p approach performed well in generating homogeneous regions for both bandwidth scenarios. Based on these findings we propose that in the context of our study, the regionalization that uses GWR coefficients, which are based on smaller bandwidths (Figure 6a) and are thus less correlated, reflects more meaningful results. The results of our study highly suggest that non-stationarity can only be reliably assessed if local coefficient estimation can be done with acceptable statistical power i.e., by defining subregions of sufficient size for which higher degrees of stationarity in the association of interest can be expected relative to the whole study area. In this article we suggested the combination of GWR and max-p regionalization (Figure 7) as an appropriate way to systematically identify such homogeneous regions of critical size.

We contend that the use of our proposed framework for assessing the need for subregional versus a country-wide disease model provides a sound basis for interpreting the observations made. For example, in both bandwidth scenarios (0.025 and 0.1) in Figure 6a and b we

found one large region (region A) in northwestern Brazil where illiteracy tested insignificant in the regional-scale models although it was highly significant in the national-scale model (Table 2). This relates to the early, unexpected observation made in the bivariate LISA maps in which clusters of low mortality rates and high percentages of illiterate population are spatially associated (Figure 3b). The same observation was made for region B in Figure 6a. Interestingly, this area (regions A and B in Figure 6a) is subject to known underreported infant mortality (Castro and Simões, 2010). Consequently, a possible reason for the creation of these max-p regions and the weak statistical association between mortality and illiteracy in this region could be poor data quality in administrative databases (i.e., risk factor or mortality data). This observation could guide investigations of data and model uncertainties to develop appropriate data correction methods (Yotthanoo and Choonpradub, 2010) that would improve statistical analysis. In other cases such regionally changing statistical associations could be driven by truly existing changes in putative associations as a regional phenomenon (possibly caused by changing environmental, climatic or socioeconomic conditions).

Our approach and suggested framework is not without limitations. More research will be needed to further test and understand the influence of the size constraint (floor constraint) in max-p on the resulting model performances. This parameter determines the geographic scale at which regional models will be estimated. In our study we derived the size of 50 CMRs as an appropriate scale since the statistical power was warranted at about 80% and the max-p algorithm was able to generate homogeneous regions without enforcing CMRs into single regions. Also, we restricted predictive variables for our disease outcome to the two variables identified in an analysis of the whole of Brazil, rather than determining if variables that have predictive power might vary across subregions. However, our framework readily lends itself to testing if different variables could be recommended (i.e., whether and where they proved to be statistically significant predictors) for disease modeling in different regions, a logical continuation of this research. Such analysis would have incredible potential to improve planning efforts for the allocation of scarce resources at the regional scale and in developing policies targeting public health related questions, but also for developing more efficient sampling designs for field surveys.

Finally, we were not able to ascertain the basis for observed missing data at the municipal level (the reporting unit) when we aggregated the data to the CMRs used as the basic geographic unit of analysis in our study. Therefore, the described associations between mortality and the risk factors, household access to water supply and prevalence of illiteracy, should not be used as definitive in terms of public health assessment. We maintain that the data, as used, however, does lend itself to the purpose of our study, which was to determine the degree of variation in regional-scale model coefficients representing such risk factors.

In conclusion, this article provides a general analytical framework to assess spatial nonstationarity in statistical associations by employing a sound analytical procedure that can be transferred to different study areas as well as to different disease outcomes. This framework should provide analysts with a methodological procedure for (a) testing the existence and impact of non-stationary associations across their study area compared to a broader-scale statistical model, (b) examining the statistical validity (significance) in using

the selected variables in individual regions, as well as (c) carrying out reliability checks to see if generated regions reflect a change in data quality or significant variations in putative associations.

Acknowledgements

This research was supported by grant number 47075 from the Bill and Melinda Gates Foundation, “Etiology, Risk Factors and Interactions of Enteric Infections and Malnutrition and the Consequences for Child Health and Development.” The research was also supported by interagency personnel agreements between the Fogarty International Center (NIH) and the University of Colorado at Boulder and Colorado State University, respectively.

References

- Akaike H. A new look at the statistical model identification. *IEEE Trans Automatic Control*. 1974; 19(6):716–723.
- Alonso WJ, Acuña-Soto R, Giglio R, Nuckols JR, Leyk S, Schuck-Paim C, Viboud C, Miller MA, McCormick BJJ. Spatio-temporal patterns of diarrhoeal mortality in Mexico. *Epidemiol Infect*. 2012; 140(1):91–99. [PubMed: 21489339]
- Anselin L. Local Indicators of Spatial Association – LISA. *Geogr Anal*. 1995; 27:93–115.
- Araújo, EM de, Costa, MCM., Oliveira, NF de, Santana, FS., Barreto, ML., Hogan, V., Araújo, TM de. Spatial distribution of mortality by homicide and social inequalities according to race/skin color in an intra-urban Brazilian space. *Rev. Bras. Epidemiol*. 2010; 13(4):549–560. [PubMed: 21180845]
- Beale L, Abellan JJ, Hodgson S, Jarup L. Methodologic issues and approaches to spatial epidemiology. *Environ Health Perspect*. 2008; 116:1105–1110. [PubMed: 18709139]
- Belsley, DA., Kuh, E., Welson, ER. Regression diagnostics. Identifying influential data and sources of collinearity. Wiley; 1980.
- Brunsdon CF, Fotheringham AS, Charlton ME. Geographically weighted regression: a method for exploring spatial nonstationarity. *Geogr Anal*. 1996; 28:281–298.
- Carrel M, Escamilla V, Messina J, Giebultowicz S, Winston J, Yunus M, Streatfield PK, Emch M. Diarrheal disease risk in rural Bangladesh decreases as tubewell density increases: a zero-inflated and geographically weighted analysis. *Int J Health Geogr*. 2011; 10:41. [PubMed: 21676249]
- Castro MC, Simões CC. Spatio-temporal trends of infant mortality in Brazil. *Proceedings of the Annual Meeting. Population Association of America (PAA)*. 2010
- Chaikaew N, Tripathi NK, Souris M. Exploring spatial patterns and hotspots of diarrhea in Chiang Mai, Thailand. *Int J Health Geogr*. 2009; 8:36. [PubMed: 19552822]
- Cheng E, Atkinson PM, Shahani AK. Elucidating the spatially varying relation between cervical cancer and socio-economic conditions in England. *Int J Health Geogr*. 2011; 10:51. [PubMed: 21943079]
- Cho S-H, Lambert DM, Chen Z. Geographically weighted regression bandwidth selection and spatial autocorrelation: an empirical example using Chinese agriculture data. *Appl Econ Lett*. 2010; 17(8):767–772.
- Cleveland, WS., Grosse, E., Shyu, WM. Local regression models. In: Chambers, JM., Hastie, TJ., editors. *Statistical Models in S*. Pacific Grove: Wadsworth & Brooks; 1991. p. 309-376.
- Cohen, J. Statistical power analysis for the behavioral sciences. Hillsdale, NJ: Lawrence Erlbaum Associates; 1988.
- Cohen J. Quantitative methods in psychology: A power primer. *Psychological Bulletin*. 1992; 112(1): 155–159. [PubMed: 19565683]
- Duque, JC., Anselin, L., Rey, SJ. The Max-P-Region problem. San Diego: Regional Analysis Laboratory, San Diego State University; 2007.
- Edwards KL, Clarke GP, Ransley JK, Cade J. The neighbourhood matters: studying exposures relevant to childhood obesity and the policy implications in Leeds UK. *J Epidemiol Commun Health*. 2010; 64:194–201.

- Elliott P, Wartenberg D. Spatial Epidemiology: Current Approaches and Future Challenges. *Environ Health Perspect.* 2004; 112(9):998–1006. [PubMed: 15198920]
- Farber S, Páez A. A systematic investigation of cross-validation in GWR model estimation: empirical analysis and Monte Carlo simulations. *J Geogr Sys.* 2007; 9:371–396.
- Fotheringham, AS., Brunsdon, C., Charlton, M. *Geographically Weighted Regression: The Analysis of Spatially Varying Relationships.* Wiley; 2002.
- Griffith D. Spatial-filtering-based contributions to a critique of geographically weighted regression (GWR). *Environ Planning A.* 2008; 40(11):2751–2769.
- Hastie T, Tibshirani R. Varying-Coefficient Models. *J Royal Stat Soc, Series B.* 1993; 55(4):757–796.
- Jepsen MR, Simonsen J, Ethelberg JS. Spatio-temporal cluster analysis of the incidence of *Campylobacter* cases and patients with general diarrhea in a Danish county, 1995–2004. *Int J Health Geogr.* 2009; 8:11. [PubMed: 19228436]
- Kelly-Hope LA, Alonso WJ, Thiem VD, Anh DD, Canh do G, Lee H, Smith DL, Miller MA. Geographical distribution and risk factors associated with enteric diseases in Vietnam. *Am J Trop Med Hyg.* 2007; 76:706–712. [PubMed: 17426175]
- Leyk S, McCormick BJJ, Nuckols JR. Effects of varying temporal scale on spatial models of mortality patterns attributed to pediatric diarrhea. *Spat and Spatio-temporal Epidemiol.* 2011; 2:91–101.
- Leyk S, Maclaurin GJ, Hunter LM, Nawrotzki R, Twine W, Collinson M, Erasmus B. Spatially and Temporally Varying Associations between Temporary Outmigration and Natural Resource Availability in Resource-Dependent Rural Communities in South Africa: A Modeling Framework. *Appl Geogr.* (in press).
- Lin C-H, Wen T-H. Using geographically weighted regression (GWR) to explore spatial varying relationships of immature mosquitoes and human densities with the incidence of dengue. *Int J Environ Res Public Health.* 2011; 8:2798–2815. [PubMed: 21845159]
- Mandal R, St-Hilaire S, Kie JG, Derryberry D. Spatial trends of breast and prostate cancers in the United States between 2000 and 2005. *Int J Health Geogr.* 2009; 8:53. [PubMed: 19785775]
- Murtagh F. A survey of algorithms for contiguity-constrained clustering and related problems. *Computer Journal.* 1985; 28(1):82–88.
- Nakaya T, Fotheringham AS, Brunsdon C, Charlton M. Geographically weighted Poisson regression for disease association mapping. *Stat Med.* 2005; 24:2695–2717. [PubMed: 16118814]
- O'Sullivan, D., Unwin, D. *Geographic information analysis.* Wiley; 2010.
- Pande S, Keyzer MA, Arouna A, Sonneveld BG. Addressing diarrhea prevalence in the West African Middle Belt: social and geographic dimensions in a case study for Benin. *Int J Health Geogr.* 2008; 7:17. [PubMed: 18433488]
- Rey SJ, Anselin L. PySAL. A Python library of spatial analytical methods. *Rev Regional Studies.* 2007; 37(1):5–27.
- Shirabe T. A model of contiguity for spatial unit allocation. *Geogr Anal.* 2005; 37(1):2–16.
- Strina A, Cairncross S, Barreto M, Larrea C, Prado M. Childhood diarrhea and observed hygiene behavior in Salvador, Brazil. *Am J Epidemiol.* 2003; 157(11):1032–1038. [PubMed: 12777367]
- Tian N, Wilson JG, Zhan FB. Spatial association of racial/ethnic disparities between late-stage diagnosis and mortality for female breast cancer: where to intervene? *Int J Health Geogr.* 2011; 10:24. [PubMed: 21463525]
- Wheeler D, Tiefelsdorf M. Multicollinearity and correlation among local regression coefficients in geographically weighted regression. *J Geogr Sys.* 2005; 7:161–187.
- Wheeler DC. Diagnostic tools and a remedial method for collinearity in geographically weighted regression. *Environ Planning A.* 2007; 39(10):2464–2481.
- Wheeler, DC., Páez, A. *Geographically Weighted Regression.* In: Fischer, MM., Getis, A., editors. *Handbook of Applied Spatial Analysis: Software Tools, Methods and Applications.* 2010. p. 461–486.
- Wimberly MC, Yabsley MJ, Baer AB, Dugan VG, Davidson WR. Spatial heterogeneity of climate and land cover constraints on distributions of tick-borne pathogens. *Global Ecol Biogeogr.* 2008; 17:189–202.

Yotthanoo S, Choonpradub C. A statistical method for estimating under-reported incidence rates with application to child Diarrhea in Thai Provinces bordering Cambodia. *Southeast Asian J Trop Med Public Health*. 2010; 41(1):203–214. [PubMed: 20578500]

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Figure 1.

Spatial and statistical distributions of mortality rates attributed to intestinal infectious disease, children under age 5 in Census Micro Regions of Brazil, year 2000.

Figure 2.

(a) Spatial distribution of national-scale Poisson model residuals using the two selected risk factors (*illit* and *wtr_h*); (b) Statistically significant local clusters of model residuals based on LISA (HH –high values; LL – low values; HL and LH – outliers).

Figure 3. Statistically significant ($p < 0.05$) spatial clusters from a bivariate LISA analysis: (a) using pediatric mortality rates attributed to intestinal infectious disease and household access to water (*wtr_h*) and (b) the same mortality variable and percent illiterate population (*illit*) (HH – high values; LL – low values; HL and LH – outliers; first letter indicates mortality, second one the risk factor).

Figure 4. Classification of the GWR coefficients for illiteracy (*illit*) using increasing proportions of CMRs (from 0.025 to 0.25) as adaptive bandwidths. Light red = Min; Dark red = Max. Coefficient values defining the class breaks are in Table 1.

Figure 5. Classification of the GWR coefficients for household water access (wtr_h) using increasing proportions of CMRs (from 0.025 to 0.25) as adaptive bandwidths. Light red = Min; Dark red = Max. Coefficient values defining the class breaks are in Table 1.

Figure 6.

Resulting max-p regions (floor constraint = 50 CMRs) based on GWR coefficients of the two risk factors, *illit* and *wtr_h* and for two bandwidth scenarios (a) 0.025 and (b) 0.1.

Colors symbolize inclusion of a CMR in a region. Region labels and colors in both maps correspond roughly in location; labels refer to rows in Table 2.

Figure 7.

Proposed analytical framework for assessing spatial non-stationarity in disease-risk factor associations across a specified study area (case study: Brazil).

Table 1

Class break values underlying the GWR coefficient surfaces in Figures 4 and 5.

Variable (risk factor)	Adaptive bandwidth	Minimum (light red in Figures 4 & 5)	1 st Quartile Q25	Median	3 rd Quartile Q75	Maximum (dark red in Figures 4 & 5)
illit ¹	0.025	-0.0641	0.0199	0.0421	0.0689	0.2795
	0.050	-0.0104	0.0295	0.0430	0.0689	0.2073
	0.075	0.0139	0.0353	0.0459	0.0664	0.1538
	0.100	0.0199	0.0362	0.0472	0.0711	0.1238
	0.125	0.0230	0.0371	0.0473	0.0700	0.1079
	0.150	0.0247	0.0381	0.0468	0.0617	0.0974
	0.175	0.0261	0.0382	0.0455	0.0543	0.0897
	0.200	0.0270	0.0382	0.0443	0.0512	0.0822
	0.225	0.0275	0.0380	0.0435	0.0494	0.0760
	0.250	0.0279	0.0380	0.0426	0.0477	0.0682
wtr_h ²	0.025	-0.0000170	-0.0000040	-0.0000020	-0.0000011	0.0000008
	0.050	-0.0000105	-0.0000025	-0.0000017	-0.0000012	-0.0000002
	0.075	-0.0000041	-0.0000022	-0.0000016	-0.0000013	-0.0000004
	0.100	-0.0000028	-0.0000020	-0.0000016	-0.0000013	-0.0000006
	0.125	-0.0000025	-0.0000019	-0.0000016	-0.0000013	-0.0000007
	0.150	-0.0000024	-0.0000018	-0.0000016	-0.0000014	-0.0000008
	0.175	-0.0000023	-0.0000018	-0.0000016	-0.0000014	-0.0000009
	0.200	-0.0000023	-0.0000018	-0.0000016	-0.0000014	-0.0000010
	0.225	-0.0000022	-0.0000018	-0.0000016	-0.0000015	-0.0000011
	0.250	-0.0000022	-0.0000018	-0.0000016	-0.0000015	-0.0000012

¹ prevalence of illiteracy in the population (illit)

² household access to water supply (wtr_h)

Table 2

Results from national-scale and regional-scale (within max-p regions) models.

Adaptive bandwidth	Region ¹	Intercept ²	Coeff. wtr_h^2 β_{nat}, β_{reg}	Coeff. $illit^2$ β_{nat}, β_{reg}	CMRs ⁵	AIC _{reg} ⁴	AIC _{nat} ³	AIC _{nat} -AIC _{reg} [%]
National-scale model	National-scale model	-8.899 ***	-0.0000017 ***	0.0343 ***	557	--	5206.9	--
0.025	A	-8.06 ***	-0.0000012 **	0.0056	66	581.73	647.20	10.12
	B	-8.35 ***	-0.0000008 *	0.0124	49	559.68	637.53	12.21
	C	-8.58 ***	-0.0000029 ***	0.0357 ***	81	803.23	881.14	8.84
	D	-11.43 ***	-0.0000001	0.0929 ***	54	394.86	455.87	13.38
	E	-9.14 ***	-0.0000029 ***	0.0417 ***	111	783.73	816.34	3.99
	F	-9.38 ***	-0.0000025 ***	0.0662 ***	79	504.30	516.07	2.28
	G	-16.02 ***	-0.0000007 ***	0.3976 ***	49	723.86	911.53	20.59
	H	-7.39 ***	-0.0000111 ***	-0.0082	60	342.00	383.23	10.76
0.1	A	-7.76 ***	-0.0000018 ***	-0.0056	88	753.86	916.26	17.72
	B	-8.00 ***	-0.0000064 ***	0.0188 ***	64	617.34	673.14	8.29
	C	-8.99 ***	-0.0000012 ***	0.0424 ***	133	1262.20	1356.72	6.97
	D	-10.30 ***	-0.0000008	0.0667 ***	49	399.65	423.06	5.53
	E	-8.95 ***	-0.0000023 ***	0.0255 *	54	320.29	346.34	7.52
	F	-7.55 ***	-0.0000072 ***	-0.0049	60	382.02	418.19	8.65
	G	-11.81 ***	-0.0000010 ***	0.1781 ***	52	771.06	833.98	7.54
	H	-7.13 ***	-0.0000113 ***	-0.0239	49	264.08	281.21	6.09

¹ max-p regions based on a floor constraint of 50 CMRs

² Significance levels for access to water (wtr_h) and percent illiterate population (illit):

*** p<0.001;

** p<0.01;

* p<0.05

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

³ AIC based on running the national-scale model on regions; CMRs only

⁴ AIC of the regional-scale model;

⁵ Census Micro Regions