HHS Public Access

Author manuscript

Trans GIS. Author manuscript; available in PMC 2017 February 08.

Published in final edited form as:

Trans GIS. 2015 December ; 19(6): 877–895. doi:10.1111/tgis.12134.

Understanding the combined impacts of aggregation and spatial non-stationarity: The case of migration-environment associations in rural South Africa

Galen Maclaurin,
University of Colorado Boulder

Stefan Leyk, and
University of Colorado Boulder

Lori M. Hunter
University of Colorado Boulder

1. Introduction

The association between human migration and environmental conditions has received increasing scholarly attention in the past several years (e.g., Gray & Bilsborrow, 2012). Yet lacking precise geographic identifiers for household-level data can yield challenges for empirically modeling the migration-environment linkages. The methodological exercise presented here reveals important spatial non-stationarity and aggregation effects that may impact empirical estimates within demographic analyses. Unique socio-demographic and environmental data from rural South Africa are used to examine models of the migration-environment association under increasing aggregation starting from the household level. Although we use the migration-environment association as illustration, lessons regarding the impacts of spatial non-stationarity and aggregation can likely be more broadly applied to a variety of socio-demographic phenomena.

1.1 Background

Demographic and socio-economic data are often collected at various levels of aggregation (i.e. census unit, county, village, etc.) and the structure of aggregation can pose significant challenges for analysis and interpretation. These challenges are referred to as the modifiable areal unit problem (MAUP) (Openshaw & Taylor, 1979; Flowerdew, Geddes & Green, 2001). A primary reason for such aggregation effects is spatial autocorrelation, or pair-wise correlation between neighbors for a given characteristic, which can lead to bias in conventional statistical methods (Cliff & Ord, 1981).

As another methodological challenge, many socio-demographic processes, such as migration, operate at a relatively local scale (i.e., the individual or household) and ideally data used to examine such processes should have a spatial resolution sufficiently close to the scale of operation (Leyk et al., 2012a). Yet confidentiality concerns constrain the geographic

identifiers typically included within publicly available individual- or household-level microdata. Even if data are available at finer spatial resolution, social and demographic phenomena often exhibit spatial non-stationarity whereby model coefficients show patterns of spatial variation (Fotheringham, Carlton, & Brunsdon, 1996).

Combined, aggregation effects and non-stationarity complicate spatial analytical procedures as applied to the investigation of socio-demographic questions. The research presented here grapples with these challenges while taking migration-environment associations as our substantive focus. Using spatially explicit (GPS-measured) household-level surveillance data from the Agincourt Health and Demographic Surveillance Site (AHDSS) in a remote rural region of South Africa, we examine the implications of aggregation effects in combination with spatial non-stationarity by developing global and local migration models fit across increasing levels of simulated data granularity. Results indicate that model association between variables can be characterized by their behavior as data are aggregated. Some associations are more sensitive and lose significance at higher aggregation levels, a phenomenon we call *operational scale sensitivity*. Such sensitivity has ramifications for choice of variables, model performance and substantive interpretation when working with aggregated data.

1.1 MAUP in the Context of Migration Modeling

The delineation of aggregated units in demographic surveys is typically an administrative and non-data driven process. Therefore, any analytical procedure, spatial or aspatial, will be influenced and confounded by the aggregation's nature (Openshaw, 1983). MAUP, and its inherent zoning and scale effects, can be seen as the geographical manifestation of ecological fallacy which occurs if inferences from an aggregated analysis are assumed to pertain to individuals (Wong, 1995; Waller & Gotway, 2004). A methodological challenge logically arises when research questions target the individual-level, which consequently defines the operational scale of the process of interest, but researchers have only group-level data (Piantadosi, Byer, & Green, 1988). In this way, there exists a mismatch between the operational and analytical scales.

Focusing on demographic data, the MAUP's impact on correlation coefficients and regression models has been studied extensively at nested aggregation levels (i.e. census blocks, block groups and tracts) (Openshaw & Taylor, 1979; Wong, 2009). Yet typically only a few aggregation levels are available for analyses. In response, some researchers have generated synthetic datasets of different granularities as opposed to using observed data (Amrhein, 1995; Steel and Holt, 1996). Reynolds & Amrhein (1998) show that synthetic data allow for more control and systematic understanding of aggregation effects. Research into the MAUP's impact on regression analysis reveal that models estimated at different levels of aggregation can yield coefficient estimates that fluctuate significantly and even exhibit changes in the estimated direction of effect (Fotheringham & Wong, 1991).

More recent research has addressed effects of random aggregation using observed data (Flowerdew, Geddes & Green, 2001), analytical bias associated with group-level compared to individual-level synthetic data (Pawitan & Steel, 2006), and, specific to this research paper, scale dependence of population-environment interactions (Walsh, Crews-Meyer,

Crawford & Welsh, 2004). These studies suggest that estimated relationships between demographic and environmental variables fluctuate across different analytical scales.

In all, the effects of the MAUP on analyses of spatially aggregated demographic data are relatively well understood (Wong, 2009). Even so, there remains a gap with regard to understanding the impacts of mismatch between the operational and analytical scales when aggregating from a finer-scale phenomenon such as individual- or household-level migration. Indeed, many demographic processes are resultant of decision-making at a relatively fine scale – consider fertility decision-making as another example. As such, local contextual influences are logically of importance, and contextual data at finer spatial resolutions would support analysis at the operational scale. Unfortunately, socio-demographic individual and household level data are commonly aggregated to coarse spatial resolution due to confidentiality concerns (e.g., PUMAS with a minimum of 100,000 people). This is likely one main reason why the impact of the MAUP on migration modeling and migration-environment associations remains widely unrecognized and understudied. Putting “people into place” is challenged by lack of precise geographic identifiers (Entwisle 2005).

1.2 Spatial Non-Stationarity in Migration Models

Spatial non-stationarity within model associations is also receiving attention within population-environment modeling. Associations exhibit spatial non-stationarity when the relationship between variables is dependent on, and varies with, observation locations. The concern here is that the associations themselves spatially vary and are, therefore, not stationary (i.e., not constant). The traditional aspatial regression model, referred to as a *global* model here, is generally fit using all observations (or units of analysis). A global model returns one set of coefficient estimates and, therefore, is unable to reveal spatial non-stationarity.

In contrast, a *local* model estimates a set of coefficients for each observation based on a user-defined neighborhood of nearby observations. Local estimators have been proposed as a means of examining non-stationarity in model associations, illustrating that spatial effects can confound global regression models (Hastie & Tibshirani, 1993; Fotheringham, Brunsdon & Charlton, 2000). A number of local estimators have been developed, including varying coefficient models (Hastie & Tibshirani, 1993), local regression (Loader, 1999), and geographically weighted regression (GWR) (Fotheringham, Brunsdon & Charlton, 2002). This family of local estimators modifies the traditional regression equation by applying a spatial weights matrix (i.e., a distance decay function) to neighboring observations. A separate regression is then run for each observation.

While GWR is the most commonly used local estimator, critiques argue that the approach lacks robustness for statistical inference and should only be used for exploratory purposes (O’Sullivan & Unwin, 2010, p. 233). In addition, GWR may induce multicollinearity (Griffith, 2008; Wheeler & Tiefelsdorf, 2005) and artificial patterns of spatial heterogeneity in coefficient surfaces (Cho, Lambert & Chen, 2010) as a result of the spatial weights matrix applied to overlapping neighborhoods.

Methodological advancements continue including the development of spatial interaction models for origin and destination effects (LeSage & Llano, 2006), which extends early work by Schelling (1971) on spatial autocorrelation in neighborhood dynamics. Recent research on spatial filtering, which decomposes data into a trend component, a spatially structured component and random noise, has been applied in linear regression models (Tiefelsdorf & Griffith, 2007; Griffith & Chun, 2014). Yet questions remain for the field at-large including the most appropriate means of defining the kernel for spatial weighting (Berk, 2008; Leyk et al., 2012b). Tiefelsdorf (2000, p. 23) argues that spatial structure is a function of the strength of spatial relationships. If these relationships in the process under examination are weak, the spatial structure is trivial and therefore should not be explicit in the modeling framework. Alternatively, if these relationships are strong and a spatially explicit model is appropriate, spatial weighting can confound the impacts of spatial structure for the reasons described above.

With specific application to migration-environment associations and in response to the concerns of spatial weighting in local estimators, an alternative modeling framework has been proposed allowing for more robust analysis and diagnostics (Leyk et al., 2012a). This framework produces distributions of local coefficient estimates without the use of a spatial weights matrix by implicitly incorporating spatial structure, characterized by the existence of spatial non-stationarity of model associations. In our earlier research, we chose this method over those discussed above for its overall simplicity and lack of assumptions about the data's spatial structure.

Using this framework, traditional, aspatial linear regression models were run on local extents defined by a spatially-constrained random region permutation. This procedure was repeated hundreds of times to produce a robust set of coefficient estimates for each observation. When modeling outmigration at the household level, this local estimator provided improved model fit compared to a global model while accounting for spatial non-stationarity without the shortcomings of spatial weighting (Leyk et al., 2012a).

Consideration of spatial non-stationarity within the migration-environment association is not only of methodological interest. In many rural communities across the globe, daily life and decisions about livelihoods are closely tied to very local environmental conditions and may, therefore, actually be characterized by important spatial variation. Such variation is not revealed within global regression models. In South Africa, for example, case studies in two rural villages demonstrate that 70% of households made use of non-timber forest products, such as fuelwood, wild fruit, and edible herbs, during times of shortage and crisis (Paumgarten & Shackleton, 2011). Even in rural South African villages with readily available electricity, over 90% of households use fuelwood as a primary energy source due to the cost of electricity and appliances (Twine, Moshe, Netshiluvhi & Siphugu, 2003). However, the proximate availability of these natural resources varies at regional and local scales and even for households within the same village. Consequently, local shifts in the availability of natural resources influence households differently and may result in livelihood adaptations that concomitantly exhibit spatial variation. One important livelihood adaptation is temporary or permanent migration by individual household members (Bilsborrow, 2002; McLeman & Hunter, 2010). Brought together, these variations in natural resources

availability and livelihood adaptations such as migration can yield spatial variation in the migration-environment association. Further, this spatial non-stationarity can be anticipated to interact with aggregation effects to additionally confound analyses and influence estimated coefficients.

While research on migration-environment associations has burgeoned over the past several years, few studies consider the impacts of spatial variability or data aggregation, not to mention their potential interaction and/or combined influence. Making use of detailed georeferenced (i.e., GPS measured) demographic data at the household level, this research examines precisely this issue.

We build on recent research demonstrating the importance of local migration models and non-stationarity (Leyk et al., 2012a), through examination of aggregation effects in both global and local statistical models as compared to the operational scale (i.e., the household level). Exploration of interactions between aggregation and non-stationarity, as related to migration-environment associations, also fills an important knowledge gap in the spatial sciences.

2. Data and Preprocessing

This study employs the 2007 household census conducted at the Agincourt HDSS in a rural region of northeastern South Africa. The study site is operated by the MRC/Wits Rural Public Health and Health Transitions Research Unit. The surveillance dataset consists of 9,374 geo-referenced households in 21 villages, with data representing 38,118 individuals. We conceptualize the household level as the operational scale (i.e., at which the process of interest takes place) since temporary migration tends to be a household, rather than an individual, decision in this region (Collinson, Wolff, Tollman & Kahn, 2006).

The study site is characterized by a decreasing west-east rainfall gradient resulting in substantial spatial variation in natural resource availability. Further, households within the region tend to rely heavily on proximate natural resources collected from communal landscapes both for sustenance and for raw materials to generate goods for sale (e.g., baskets, mats) (Hunter, Twine & Patterson, 2007; Twine, Moshe, Netshiluvhi & Siphugu, 2003). Finally, approximately 20 percent of residents annually engage in circular, temporary migration in which a migrant does not permanently relocate but moves between home and workplace with various regularity (Collinson et al., 2006; Collinson, Wolff, Tollman & Kahn, 2006). Detailed demographic and spatial data combined with information on local environmental conditions provide a unique modeling opportunity to study the combined effects of aggregation and spatial non-stationarity in the migration-environment association.

The models estimate the number of temporary migrants (tempmign) as the outcome. A temporary migrant is defined as 15+ year old household member who spends more than six months in a year away from home while remaining linked to the household.

We focus on the association between temporary migration and locally available natural resources. A “greenness” variable was created from MODIS-derived Normalized Difference Vegetation Index (NDVI) surfaces by first creating a 2000-meter buffer around each

household, informed by an understanding of the distance residents tend to travel to collect natural resources (Giannecchini et al., 2007). The area within village boundaries was excluded from these buffers since it is not used for natural resource collection. The sum of NDVI pixel values within the buffer (outside of villages) was divided by the number of households inside the buffer, representing an approximation of per household availability of communal natural resources.

Although our analytical focus is the migration-environment association, prior migration scholarship, in addition to our earlier modeling efforts with Agincourt data, suggest the importance of education and socio-economic status (SES) as migration correlates (Hunter et al., 2014; White & Lindstrom, 2006). As such, we also include indicators of total years of education for all household members (HHeduc) as well as SES. In rural settings, monetary and non-monetary forms of income are common, and are subject to seasonal fluctuation (Montgomery et al., 2000). Thus, an additive index is used to reflect SES which includes household assets important to local livelihoods and reflective of livelihood security. These include five major asset categories: modern assets (e.g., cell phone and automobile ownership), livestock assets, energy sources, dwelling material, and, water and sanitation (Agincourt HDSS, 2009).

Also based on prior migration scholarship (e.g. White & Lindstrom, 2006), the following are included as control variables at the household-level: female head (Boolean variable—femhead), married head (Boolean variable—marhead), proportion working (HHwork), proportion male (mascprop), dependency ratio (members over 65 years divided by members between 15–65 years—deprop), and household size (HHpop). Household size is modeled as an independent variable rather than an exposure variable (i.e., an offset).[1] No significant collinearity was found among the independent variables using the variance-inflation factor (VIF) as a diagnostic tool (Hill & Adkins, 2007).

3. Methods

Our aim is to provide a modeling framework for comparatively investigating aggregation effects on local and global regression models of outmigration. The local model (described in more detail below) has been adjusted and extended from a method developed by the authors (Leyk et al., 2012a).

As a point of reference, both the local and global models are first computed at the household level, referred to as the *baseline*. As discussed above, this is the *operational level*, the scale at which the process of interest—household temporary outmigration—predominantly operates. Using the positional measurements, household data are systematically aggregated to spatial units of increasing size. These aggregations are then used as analytical units in estimating outmigration with both global and local models. This comparative approach allows

[1]While larger households have more potential to send a migrant, the relationship is not linear (or log-linear). There is a moderate correlation between the number of temporary migrants and household size (0.58) which, although substantial, we argue does not represent an exposure variable. Use of an offset fixes the coefficient to 1 and forces a log-linear relationship with the dependent variable. Therefore we simply control for household size and estimate the regression coefficient.

examination of model behavior for aspatial global models versus local models which account for spatial non-stationarity across increasing aggregation levels.

3.1 Creating aggregated analytical units

The aggregation algorithm is based on a so-called binary partition tree scheme which recursively divides the household-level data (household locations) into subregions until each of these regions meets a size criterion defined by the number of households. Starting with the original household locations (i.e., the complete set of point features), the algorithm randomly chooses two seed points (household locations) and then groups all remaining households to one of the two seed points based on minimum (Euclidean) distance. This creates two spatially contiguous regions which are each then subdivided again into two new contiguous regions using the same procedure. This procedure is repeated until each of the resulting subregions has a number of households that is within a set range of thresholds. The lower threshold determines the minimum number of households for each region, and the upper threshold is twice this minimum, allowing the size of regions to vary in a way that is similar to an administrative unit (e.g., a rural community or a census unit). While other random regionalization algorithms are available (e.g. Rey & Anselin, 2010), this approach was developed for computational efficiency and because of the partitioning scheme's simplicity.

Nine levels of aggregation are examined: level one has regions (analytical units) with 2 to 4 households; level two has regions with 3 to 6 households, and so forth. Level nine has regions with 10 to 20 households. Aggregating to regions larger than those at level nine with this dataset would result in too few sample units on which to run the local model. Nine aggregation levels are sufficient in this study, as the main goal is to better understand how aggregation impacts the analysis in comparison to the household level (the *baseline*). Observing such fine aggregation steps is only possible if household level spatial identifiers are available.

The groups of households in the determined subregions represent the new units of analysis at each aggregation level and are spatially referenced by the region's centroid. Attributes of the households within one region are aggregated by calculating the mean for HHeduc, HHwork, deprop, mascprop, SES and Greenness, and the sum for tempmign, HHpop, femhead, and marhead. This aggregation procedure is repeated 500 times for each level (one through nine) to allow robust comparisons of the models described below.

Note that each simulation at a given aggregation level produces a unique outcome of analytical units (regions) due to the random choice of seed points for the regions. Thus the study area is partitioned differently for each simulation, and therefore locations of resulting centroids do not correspond. Accordingly, the summarized attributes for the aggregated units vary across the 500 simulations.

Sensitivity to the number of simulations was examined and we found that beyond 300 simulations, the overall results did not vary significantly. The coefficient of variation is less than 10% for variables across the nine aggregation levels, and did not change significantly from 300 to 500 simulations. Local neighborhoods created from a spatially constrained

permutation has a much smaller number of possible configurations than an unconstrained simulation, and therefore 500 simulations were considered sufficient to generate robust results within this modeling framework.

3.2 Statistical modeling

For the non-aggregated data (baseline) and for each of the 500 simulations at each aggregation level (one through nine), we estimate both global and local regression models using the same variables. This allows direct comparison of results and the corresponding residual surfaces across aggregation levels, as well as between global and local models. Our dependent variable is the count of temporary migrants per household or per spatially aggregated unit of analysis (groups of households) which follows a Poisson distribution.

The data were tested at all levels of aggregation for overdispersion using the log-likelihood ratio test implemented in R statistical software for count data by Jackman (2012). The log-likelihoods of a negative binomial model and a Poisson model were compared to test the validity of the assumption of equal conditional mean and variance for a Poisson model (Cameron & Trivedi, 1998). The log-likelihood ratio test indicated that at each aggregation level, a Poisson model was preferred over a negative binomial model. The data were also tested for zero-inflation using the Vuong likelihood ratio test (Vuong, 1989) against a Poisson model. At all levels of aggregation the test rejected the zero-inflated Poisson model for the standard Poisson model with a significance level of 0.05. Thus a Poisson Generalized Linear Model (GLM) with the standard log link function is employed in the global and as well as in the local model (described below):

$$\log(\hat{\mu}_i)=x_i'\beta$$

where $\hat{\mu}_i$ is the expected value of the dependent variable at location i, x_i is a vector of predictor variables and β is a vector of coefficient estimates. Deviance (log-likelihood) residuals of the Poisson GLM are asymptotically normal, allowing for robust analysis of error surfaces for spatial autocorrelation using Moran's I (Lin & Zhang, 2007). In a Poisson model, the deviance residual for location i is calculated as:

$$r_i^D=\left\{\begin{array}{ll}\sqrt{2(y_i\log\left(\frac{y_i}{\hat{\mu}_i}\right)-y_i+\hat{\mu}_i)}, & if\ (y_i-\hat{\mu}_i)>0;\\ -\sqrt{2(y_i\log\left(\frac{y_i}{\hat{\mu}_i}\right)-y_i+\hat{\mu}_i)}, & if\ (y_i-\hat{\mu}_i)<0;\end{array}\right\}$$

where y_i is the dependent variable and $\hat{\mu}$ is the same as above (Agresti, 2002).

We assess spatial autocorrelation of the deviance residuals from both the global and local models using a permutation test for Moran's I (Cliff & Ord, 1981). Specifically, we use a Monte Carlo test based on 999 random permutations under the asymptotic normality assumption of Poisson deviance residuals and normalized Moran's I statistics (i.e., z-scores) are reported. The implementation of a permutation test on linear regression residuals is directly applicable to deviance residuals from a GLM (Lin & Zhang, 2007). To further

investigate the spatial structure of error surfaces, local indicators of spatial association (LISA) (Anselin, 1995) are used to identify local clustering of high or low deviance residuals. Spatial autocorrelation and local clusters in model residuals often indicate systematic over- or under-prediction resulting in non-random error structures.

3.3 Global statistical models under aggregation

The global Poisson GLM is run for each of the 500 simulated partitions at each aggregation level (one through nine). Coefficient estimates, corresponding p-values, Moran's I for residuals, and the number of statistically significant LISA clusters ($p<0.05$) from residuals are recorded. Thus, for each level of aggregation, 500 sets of global model results are stored. The results are then summarized by aggregation level providing: average coefficients and their standard deviations, the proportion significant for each variable ($p<0.05$), average Moran's I and the average number of significant local clusters in model residuals derived from LISA analysis. The global model is run only once on the non-aggregated data (*baseline*).

3.4 Local statistical models under aggregation

The local modeling for temporary outmigration is based on a random region permutation approach, similar to the aggregation procedure. The algorithm randomly generates subregions and a Poisson GLM is fitted to each of these subregions thus allowing for robust estimation of existing statistical relationships at local geographic scales (extents). The same equation used in the global model is used applied here to a local neighborhood of observations, and then repeated hundreds of times. This allows direct for comparison with the global model results, providing better insight into the impacts of aggregation and spatial non-stationarity.

One important advantage of this random spatial permutation approach is model simplicity. There are no assumptions about the data's underlying spatial structure, and the model is therefore more parsimonious than commonly used spatial models (i.e. GWR, spatial lag or spatial error models). Thus spatial autocorrelation of model associations is implicit rather than assumed and structured into the algorithm. Furthermore, the GLM framework allows for extensive model diagnostics.

At the baseline and for each level of aggregation (and herein for each of the 500 simulated partitions), spatially contiguous subregions were randomly generated using the same binary partition tree algorithm described above for aggregation although with a different intention. We now use the resulting subregions as geographic extents across which local models are fit on the units of analysis (i.e., households at the baseline level and aggregated units (groups of households) at all other levels). To ensure sufficiently large populations similar in size at each aggregation level, the subregions generated for local modeling are constrained to between 100 and 200 analytical units. These thresholds are based on preliminary analysis to minimize prediction error and optimize the Akaike Information Criterion (AIC) while maintaining sufficient degrees of freedom for statistical inference. To illustrate the local modeling procedure, when analyzing the baseline data, units (households) are randomly partitioned into spatially contiguous regions containing between 100 and 200 units. A

Poisson GLM is run on the units (households) within each region. The coefficient estimates, their corresponding p-values and model residuals are stored for all units in each region. This procedure is repeated 500 times, each time randomly generating a different partition (permutation) of regions for local modeling. Then for each unit, the mean is calculated across all 500 permutations for the coefficient estimates of each independent variable and for the model residuals. The proportion significant for each variable ($p<0.05$) is calculated and recorded for each unit across all permutations. Thus the results of the 500 local model runs are summarized for each unit (household). Testing for sensitivity to the number of simulations of the local model found that after 350 permutations, variation in results was not statistically significant. This local modeling procedure is then repeated at each of the nine levels of aggregation. However, in each case after the baseline, the units of analysis are not households but the aggregated units (i.e., groups of households) as described above. Thus subregions created for local modeling are constrained to 100–200 units of analysis, which, for instance, are aggregates of 10–20 households for aggregation level nine. As such, the number of subregions created across the study area for local modeling decreases with increasingly aggregated units of analysis. The local GLM permutation model applied at the household level is now conducted on the corresponding aggregated units of analysis for each of the 500 simulated aggregations at each of the nine aggregation levels. This framework results in 2.25 million local model permutations for levels one through nine (plus 500 local model runs for the baseline), emphasizing the importance of high degrees of efficiency.

Our overall objective is to contrast results from the global and local models at each level of aggregation. This detailed comparison will allow for improved understanding of the impacts of aggregation and spatial non-stationarity as they manifest in migration-environment associations as well as with other important migration predictors.

4. Results

4.1 Global model results

As shown in Table 1, the coefficient estimates for Greenness remain relatively stable, positive, and are statistically significant across all levels—indicating that in a global model proximate natural resources remain a strong predictor of temporary migration and stable under various levels of aggregation. This is in line with prior work in the Agincourt HDSS and elsewhere suggesting assets in the form of natural resources provide a foundation from which livelihood migration may occur – the “natural capital hypothesis” (Gray, 2009; Hunter et al., 2014).

Similar to the migration-environment association, the coefficients for HHeduc are, on average, relatively consistent and highly significant across simulations for each aggregation level. The positive value is also in line with substantive migration research linking higher education to higher migration probabilities (White & Lindstrom, 2006).

However, the mean coefficient values for SES decrease substantially with increasing aggregation and, on average, are not statistically significant ($p<0.05$) at level three or beyond. These shifting results indicate that SES is a significant predictor only at local scales, close to the operational household level. Substantively, SES exhibits a positive association

with temporary outmigration consistent with research in some regions that demonstrate the necessity of household assets to support migration's costs (e.g., Gray & Mueller, 2012). Interestingly, just as the variable loses explanatory power at increasing aggregation within the Agincourt HDSS, the positive substantive association is also not consistent across regional scales in rural South Africa.

In all, Greenness and HHeduc are stable under aggregation and do not suffer from 'operational scale sensitivity' in the global model. In contrast, the observed instability of SES raises concern when making inferences from statistical models based on aggregated data, and could lead researchers to omit variables from their analysis which are indeed important at the operational level. This phenomenon can be viewed as related to MAUP as a geographical manifestation of ecological fallacy (Waller & Gotway, 2004).

4.2 Local model results

To visualize local model results across all 500 simulations at each aggregation level, the point vector features (household locations at baseline and centroids of aggregated analytical units at all other levels) were converted to a raster representation with 30m resolution. Such a data reduction and visualization strategy was necessary since the modeling process resulted in extremely large datasets. For example, running 500 simulations at level one alone resulted in 1.6 million points. For each independent variable, mean coefficients and proportions significant were calculated from all points inside a given raster cell. Hence this conversion process results in surfaces of regression coefficients and proportions significant for each variable. Coefficient estimates and proportions significant of Greenness, HHeduc, and SES are shown for the baseline level and three aggregation levels in Figs. 1 through 6 as they represent the trends across all nine aggregation levels.

The Greenness coefficient surfaces show the highest spatial variation at the baseline level (Fig. 1), yet the most rapid smoothing process under aggregation. This leads to a homogenized coefficient surface by level five, which changes little by level nine. This trend, however, is not mirrored by the surfaces of the proportion significant (Fig. 2), which present a smoothing process that is more stable than the other two variables discussed below. Specifically, note that areas of high proportion significant remain while coefficients approach zero and coefficient surfaces become rather spatially homogenous overall.

The surfaces of HHeduc coefficient estimates (Fig. 3) and the corresponding proportions significant (Fig. 4) are highly correlated as can be seen in the spatial distributions (i.e. highly positive and highly negative local model coefficients correspond spatially with high proportions significant). Pockets of high coefficient values and high proportions significant indicate a considerable degree of spatial variation across the study area in model performance indicative of spatial non-stationarity in statistical associations. In other words, the migration-HHeduc association varies substantially across the Agincourt HDSS.

In addition, the results for HHeduc reveal significant changes with increasing aggregation in the surfaces reflecting coefficient estimates and proportions significant. Changes in the coefficient sign occur in a few regions between the baseline and level one (see Fig. 3). A general smoothing effect is apparent, in addition to a diminished pattern of spatial non-

stationarity in the coefficient surfaces with increasing aggregation (see Fig. 3). This leads to a rather regional phenomenon at level nine (bottom right panel). While this phenomenon is expected from the local model under aggregation, there is an important substantive shift with areas of high positive significant coefficients at the baseline changing to areas of negative non-significant coefficients at level nine. This is accompanied by an increasingly noticeable north-west to south-east gradient in coefficient values and proportions significant with increasing aggregation. Spatial non-stationarity of model associations is greatly reduced in the aggregation process, leading rapidly towards a homogenous, global trend.

For SES, the patterns are similar to HHeduc with a few important differences. Very local pockets of positive high-valued coefficients (Fig. 5) overlap with the highest proportions significant (Fig. 6), as seen with HHeduc. However, the smoothing process of spatial non-stationarity in the local model associations is stronger by level nine (see Fig. 5). Both coefficient and proportion significant surfaces show a stronger smoothing effect than HHeduc indicating less stability of the surface at higher levels of aggregation. In contrast to the surfaces related to HHeduc, SES shows only a weak gradient at level nine across the study site. The decrease in the range of coefficient estimates and the low proportions significant indicate that SES is less stable and has lower explanatory power across aggregation levels compared to HHeduc. This is in line with the global results.

4.3 Residual analysis

Spatial autocorrelation of model residuals measured by the Moran's I z-scores is systematically lower for local models as compared to global models (Table 2). A trend of increasing Moran's I is seen in both cases suggesting that spatial autocorrelation is more severe in the error surfaces at higher aggregation levels. The Moran's I statistic was significant ($p<0.05$) for all model runs on all aggregation levels for both global and local models.

The trend of spatial structure is similar between global and local model residuals across aggregation levels, although with systematically lower numbers of statistically significant LISA clusters for the local model (Table 2). This indicates that local models show lower degrees of local clustering in their residuals, which has been explored more extensively with this dataset in recent research (Leyk et. al., 2012a).

5. Discussion and Concluding Remarks

The association between human migration and environmental conditions has received increasing scholarly attention in the past several years (e.g., Gray & Bilsborrow, 2012). Much of this attention has been driven by concern with the potential effects of climate change on livelihood viability and migration as a possible adaptation (e.g., Adamo & Izazola, 2010). The research presented here offers methodological insight from a geographical perspective for researchers examining these migration-environment linkages. Specifically, the methodological exercise reveals important spatial non-stationarity and aggregation effects that may influence and bias statistical inference.

The global models presented here shed light on aggregation effects in statistical migration models if spatial non-stationarity is not considered. Local natural resource availability (Greenness) remains an important positive predictor of temporary migration even at very high levels of aggregation, as does the household educational level (HHeduc). Yet socioeconomic status (SES) reveals significant predictive power only at scales close to the operational scale, the household. In this way, empirical research examining the migration-environment association at higher levels of aggregation may miss important relationships that tend to reveal only closer to the operational scale.

Yet, once spatial non-stationarity is accounted for through the use of local models, the associations with migration evidenced by natural resources and household education become less similar. In general, this indicates that natural resource availability is more consistently linked with outmigration across the study site, while the association between household education and migration exhibits more geographic variability. As represented in terms of predicted spatial coefficient surfaces, the Greenness variable presents a much stronger smoothing process under aggregation in which local model associations rapidly approach a relatively homogenous surface. This is likely a response to the underlying spatial structure of the non-aggregated variables – the Greenness variable is derived using a distance measure and is therefore highly spatially autocorrelated by construction; neighbors have very similar availability of natural resources. On the other hand, the HHeduc variable presents a heterogeneous spatial pattern, while the SES variable falls somewhere in between (higher spatial autocorrelation than HHeduc yet much lower than Greenness). This is confirmed by the Moran's I values for the raw non-aggregated variables of 0.16, 0.18 and 0.99 for HHeduc, SES and Greenness, respectively.

The results from the local models allow for more in-depth interpretations: increasing levels of aggregation reduce the effects of spatial non-stationarity in the local model relationships. While this is an expected outcome, our study provides important quantitative evidence of a decrease in local variation which can result in an overall smoothing of coefficient surfaces with either low (e.g., SES) or higher proportions significant (e.g., Greenness) or in a rather regional gradient of coefficient values and their proportions significant (e.g., HHeduc).

The MAUP's aggregation effect has been extensively examined in the spatial analysis of demographic data, but mostly at higher levels of aggregation. Researchers rarely have the opportunity to compare results from different aggregation levels to the operational scale. In this study, household level demographic survey data containing geographic coordinates from a rural area in South Africa are used to examine effects of aggregation on models of temporary outmigration in comparison to the operational scale of the process of interest, the household-level. Moreover this study examined aggregation effects within two different model frameworks, a global and a local approach. By including a local model it was possible to evaluate interactions between aggregation effects and local migration-related associations while accounting for inherent spatial non-stationarity. Such interactions have not been investigated to-date.

Overall, the results reveal that specific migration predictors can show very different behavior in both global and local models under aggregation possibly due to their underlying spatial

distribution at the household level, as well as the way they are constructed or derived. These observations emphasize that model relationships can be influenced by "operational scale sensitivity." Such sensitivity may influence the choice of variables to be included, as well as a model's performance, estimated coefficients and substantive interpretation.

Future research will focus on the substantive dimension of these findings which could lead to the formulation of more general interpretations regarding the role of different predictive variables in migration models at various spatial scales. In the longer term, this modeling framework will be tested for other demographic processes of interest in order to examine its more general usability.

Acknowledgments

We acknowledge the research contributions of Raphael Nawrotzki of the University of Colorado Boulder and Wayne Twine, Mark Collinson and Barend Erasmus of the University of the Witwatersrand, South Africa. Supported by NIH R03 HD061428, "Environmental Variability, Migration, and Rural Livelihoods." The work has also benefited from the NICHD-funded University of Colorado Population Center (grant R21 HD51146) for research, administrative, and computing support. This work was also indirectly supported by the Wellcome Trust (grant 085477/Z/08/Z) through its support of the Agincourt Health and Demographic Surveillance System. The content is solely the responsibility of the authors and does not necessarily represent the official views of the CUPC, Wellcome Trust, NIH, or NICHD.

References

Adamo S, Izazola H. Human Migration and the Environment. Population and Environment. 2010; 32:105–108.

Agincourt HDSS. Core Data Dictionary. Johannesburg, South Africa: University of the Witwatersrand; 2009. Available at http://www.agincourt.co.za/wp-content/uploads/2012/10/ADSS_1in10_Dictionary.pdf [accessed 24 September 2014]

Agresti, A. Categorical data analysis. 2. Hoboken, New Jersey: John Wiley & Sons; 2002.

Amrhein C. Searching for the elusive aggregation effect: evidence from statistical simulations. Environment and planning A. 1995; 27(1):105–119.

Anselin L. Local indicators of spatial association – LISA. Geographical Analysis. 1995; 27(2):93–115.

Berk, RA. Statistical learning from a regression perspective. Springer; 2008.

Bilsborrow R. Migration, population change, and the rural environment. Environmental Change and Security Project Report. 2002; 8(1):69–84.

Cameron, AC., Trivedi, PK. Regression analysis of count data. New York: Cambridge University Press; 1998.

Cho SH, Lambert D, Chen Z. Geographically weighted regression bandwidth selection and spatial autocorrelation: an empirical example using Chinese agriculture data. Applied Economics Letters. 2010; 17(8):767–772.

Cliff, AD., Ord, JK. Spatial processes: models & applications. Vol. 44. London: Pion; 1981.

Collinson, M., Tollman, S., Kahn, K., Clark, S., Garenne, M. Highly prevalent circular migration: households, mobility and economic status in rural South Africa. In: Tienda, M.Findley, S.Tollman, S., Preston-Whyte, E., editors. Africans on the Move: Migration in Comparative Perspective. Johannesburg, South Africa: Wits University Press; 2006. p. 194-216.

Collinson M, Wolff B, Tollman S, Kahn K. Trends in internal labour migration from rural Limpopo Province: Male risk behaviour, and implications for the spread of HIV/AIDS in rural South Africa. Journal of Ethnic and Migration Studies. 2006; 32(4):633–648. [PubMed: 20396611]

Entwisle B. Putting people into place. Demography. 2007; 44(4):687–703. [PubMed: 18232206]

Fischer MM, Griffith DA. Modeling Spatial Autocorrelation In Spatial Interaction Data: An Application To Patent Citation Data In The European Union. Journal of Regional Science. 2008; 48(5):969–989.

Flowerdew, R., Geddes, A., Green, M. Behaviour of Regression Models under Random Aggregation. In: Tate, N., Atkinson, P., editors. Modelling scale in geographical information science. Chichester: Wiley; 2001. p. 89-104.

Fotheringham A, Wong D. The Modifiable Areal Unit Problem in Multivariate Statistical Analysis. Environment and Planning A. 1991; 23(7):1025–1044.

Fotheringham, A., Brunsdon, C., Charlton, M. Quantitative Geography: perspectives on spatial data analysis. London: Sage; 2000.

Fotheringham, A., Brunsdon, C., Charlton, M. Geographically Weighted Regression: The Analysis of Spatially Varying Relationships. New York: Wiley; 2002.

Fotheringham A, Charlton M, Brunsdon C. The geography of parameter space: an investigation of spatial non-stationarity. International Journal of Geographical Information Systems. 1996; 10(5): 605–627.

Giannecchini M, Twine W, Vogel C. Land-cover change and human environment interactions in a rural cultural landscape in South Africa. The Geographical Journal. 2007; 173:26–42.

Gray C. Environment, Land, and Rural Out-migration in the Southern Ecuadorian Andes. World Development. 2009; 37(2):457–468.

Gray C, Bilsborrow R. Environmental Influences on Human Migration in Rural Ecuador. Demography. 2013; 50(4):1217–1241. [PubMed: 23319207]

Gray C, Mueller V. Natural disasters and population mobility in Bangladesh. Proceedings of the National Academies of Sciences. 2012; 109(16):6000–6005.

Griffith D. Spatial-filtering-based contributions to a critique of geographically weighted regression (GWR). Environment and Planning A. 2008; 40(11):2751–2769.

Griffith, D., Chun, Y. Handbook of Regional Science. Springer; Berlin Heidelberg: 2014. Spatial autocorrelation and spatial filtering; p. 1477-1507.

Hastie T, Tibshirani R. Varying-coefficient models. Journal of the Royal Statistical Society: Series B. 1993; 55(4):757–796.

Hill, R., Adkins, L. Collinearity. In: Baltagi, B., editor. A companion to theoretical econometrics. Oxford: Basil Blackwell; 2007. p. 256-278.

Hunter L, Nawrotzki R, Leyk S, Maclaurin G, Twine W, Collinson M, Erasmus B. Rural Outmigration, Natural Capital, and Livelihoods in Rural South Africa. Population Space and Place. 2014; 20:402–420.

Hunter L, Twine W, Patterson L. 'Locusts Are Now Our Beef': Adult Mortality and Household Dietary Use of Local Environmental Resources. Scandinavian Journal of Public Health. 2007; 25(69):165–174.

Jackman, S. pscl: Classes and Methods for R Developed in the Political Science Computational Laboratory. Stanford University. Department of Political Science, Stanford University; Stanford, California: 2012. R package version 1.04.4. http://pscl.stanford.edu/

LeSage, JP., Llano, C. A spatial interaction model with spatially structured origin and destination effects. SSRN; 2006 Jul. http://ssrn.com/abstract924603

Leyk S, Maclaurin G, Hunter L, 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. Applied Geography. 2012a; 34:559–568. [PubMed: 23008525]

Leyk S, Norlund PU, Nuckols JR. Robust Assessment of Spatial Non-Stationarity in Model Associations Related to Pediatric Mortality due to Diarrheal Disease in Brazil. Spatial and Spatio-temporal Epidemiology. 2012b; 3:95–105. [PubMed: 22682436]

Lin G, Zhang T. Loglinear residual tests of Moran's I autocorrelation and their applications to Kentucky breast cancer data. Geographical Analysis. 2007; 39(3):293–310.

Loader, C. Local Regression and Likelihood. NY: Springer-Verlag; 1999.

McLeman R, Hunter L. Migration in the Context of Vulnerability and Adaptation to Climate Change: Insights from Analogues. Wiley Interdisciplinary Reviews: Climate Change. 2010; 1(3):450–461. [PubMed: 22022342]

Montgomery MR, Gragnolati M, Burke KA, Paredes E. Measuring living standards with proxy variables. Demography. 2000; 37(2):155–174. [PubMed: 10836174]

O'Sullivan, D., Unwin, D. Geographic information analysis. 2. Hoboken: Wiley; 2010.

Openshaw, S., Taylor, P. A million or so correlation coefficients: three experiments on the modifiable areal unit problem. In: Wrigley, N., editor. Statistical Applications in the Spatial Sciences. London: Pion; 1979. p. 127-144.

Openshaw, S. The modifiable areal unit problem. Vol. 38. Norwick: Geo Books; 1983.

Paumgarten F, Shackleton C. The role of non-timber forest products in household coping strategies in South Africa: The influence of household wealth and gender. Population and Environment. 2011; 33(1):108–131.

Pawitan G, Steel DG. Exploring a relationship between aggregate and individual levels spatial data through semivariogram models. Geographical Analysis. 2006; 38(3):310–325.

Piantadosi S, Byar DP, Green SB. The ecological fallacy. American Journal of Epidemiology. 1988; 127(5):893–904. [PubMed: 3282433]

Rey, SJ., Anselin, L. PySAL: A Python library of spatial analytical methods. In: Fischer, M., Getis, A., editors. Handbook of applied spatial analysis. Berlin: Springer; 2010. p. 175-193.

Reynolds, H., Amrhein, C. Some effects of spatial aggregation on multivariate regression parameters. In: Griffith, D.Amrhein, C., Huriot, JM., editors. Econometric Advances in Spatial Modelling and Methodology. Springer; US: 1998. p. 85-106.

Schelling TC. Dynamic models of segregation. Journal of mathematical sociology. 1971; 1(2):143–186.

Steel D, Holt D. Rules for random aggregation. Environment and Planning A. 1996; 28(6):957–978.

Tiefelsdorf, M. Modelling Spatial Processes: The Identification and Analysis of Spatial Relationships in Regression Residuals by Means of Moran's I. Berlin: Springer Verlag; 2000. Lecture Notes in Earth Sciences

Tiefelsdorf M, Griffith DA. Semiparametric filtering of spatial autocorrelation: the eigenvector approach. Environment and Planning A. 2007; 39(5):1193.

Twine W, Moshe D, Netshiluvhi T, Siphugu M. Consumption and direct-use values of savanna bio-resources used by rural households in Mametja, a semi-arid area of Limpopo province, South Africa. South African Journal of Science. 2003; 99:467–473.

Waller, L., Gotway, C. Applied Spatial Statistics for Public Health Data. Hoboken, NJ: Wiley; 2004.

Walsh, S., Crews-Meyer, K., Crawford, T., Welsh, W. Population and Environment Interactions: Spatial Considerations in Landscape Characterization and Modeling. In: Sheppard, E., McMaster, R., editors. Scale and Geographic Inqiuiry. Malden, MA: Blackwell; 2004. p. 41-65.

Wheeler D, Tiefelsdorf M. Multicollinearity and correlation among local regression coefficients in geographically weighted regression. Journal of Geographical Systems. 2005; 7(2):161–187.

White, M., Lindstrom, D. Internal Migration. In: Poston, D., Micklin, M., editors. Handbook of Population. New York: Springer; 2006. p. 311-343.

Wong, D. Aggregation Effects in Geo-referenced Data. In: Arlinghaus, S., editor. Practical handbook of spatial statistics. Boca Raton: CRC Press; 1995. p. 83-106.

Wong, D. The modifiable areal unit problem (MAUP). In: Fotheringham, A., Rogerson, P., editors. The SAGE handbook of spatial analysis. Los Angeles: SAGE; 2009. p. 105-123.

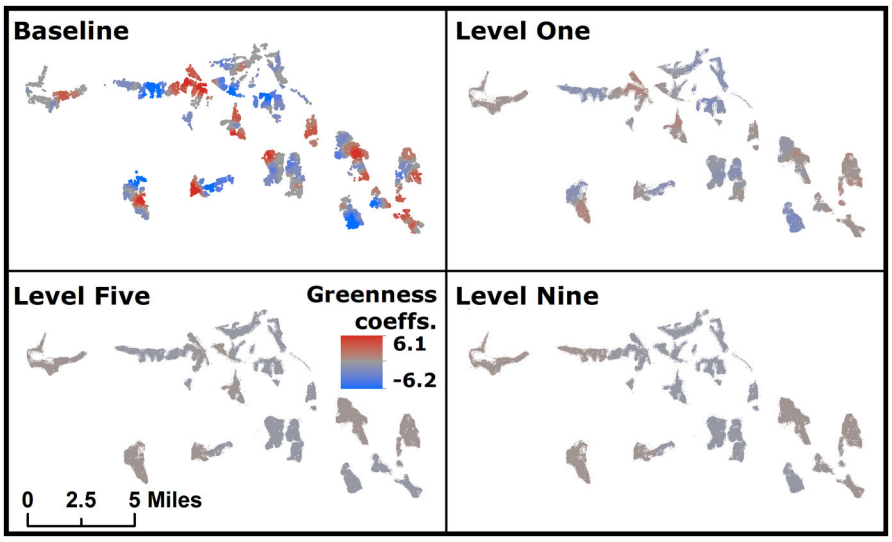

Figure 1.
Spatial distribution of coefficient estimates for Greenness from local models computed for household level (baseline) and nine aggregation levels (levels one, five and nine shown).

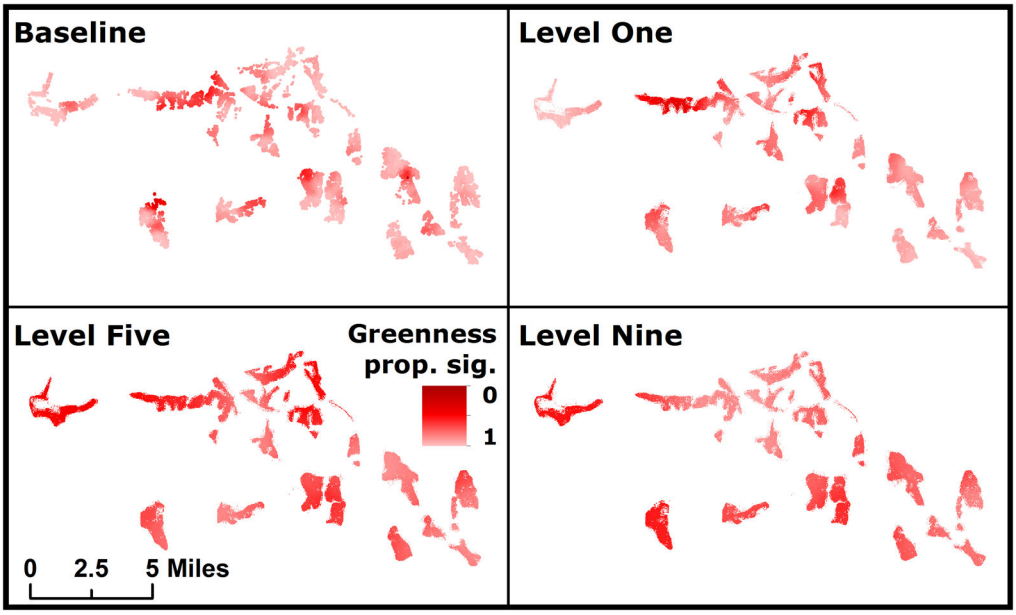

Figure 2.

Spatial distributions of the proportions significant of the coefficient estimates from local models for the Greenness variable across different levels of aggregation as shown in Figure 1.

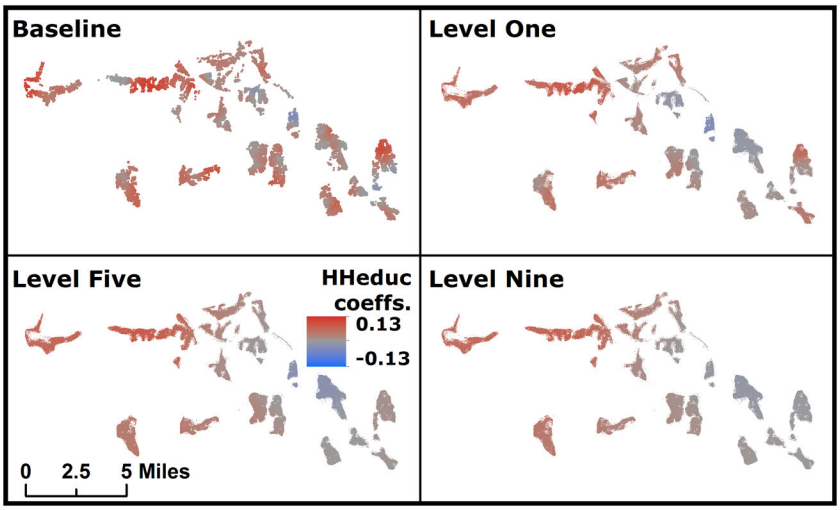

Figure 3.
Spatial distribution of coefficient estimates for household education (HHeduc) from local models computed for household level (baseline) and nine aggregation levels (levels one, five and nine shown).

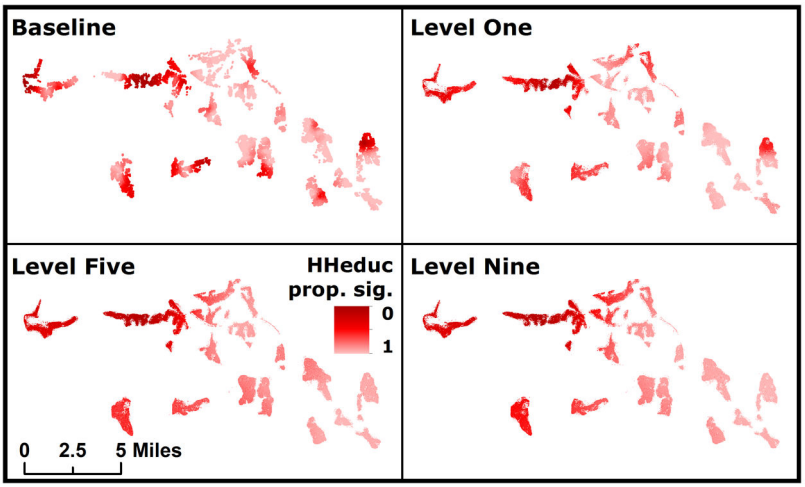

Figure 4.

Spatial distributions of the proportions significant of the coefficient estimates from local models for the HHeduc variable across different levels of aggregation as shown in Figure 3.

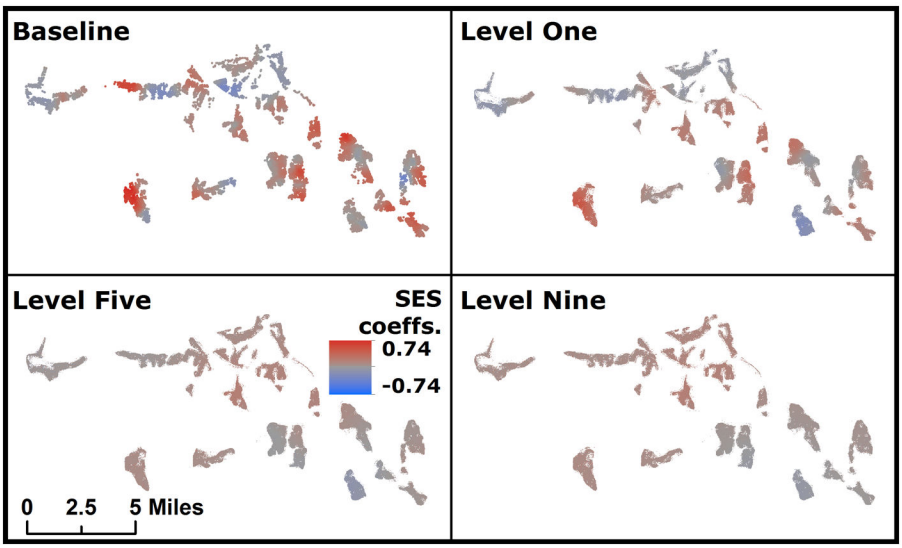

Figure 5.
Spatial distribution of coefficient estimates for socio-economic status (SES) from local models computed for household level (baseline) and nine aggregation levels (levels one, five and nine shown).

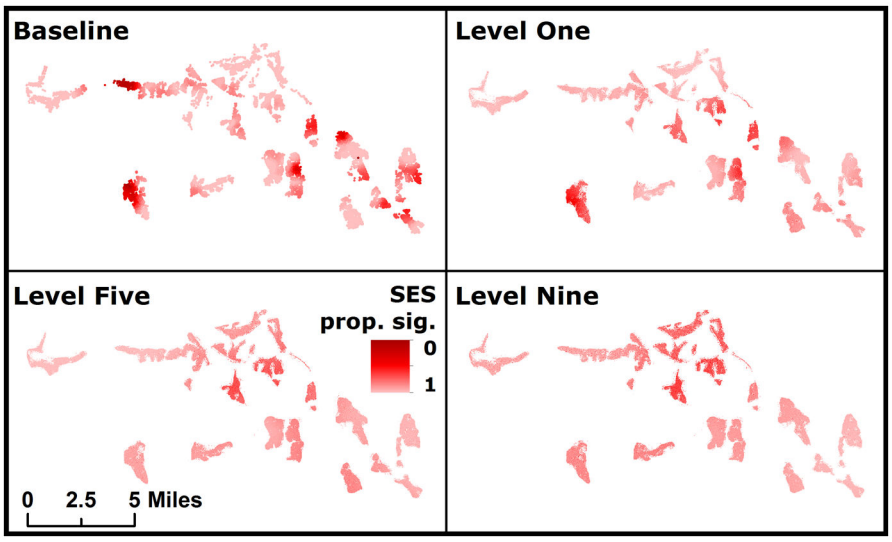

Figure 6.

Spatial distributions of the proportions significant of the coefficient estimates from local models for the SES variable across different levels of aggregation as shown in Figure 5.

Trans GIS. Author manuscript; available in PMC 2017 February 08.

Table 1

Global model coefficients and p-values (calculated as actual estimations (in parentheses) for the baseline level and the mean and standard deviation across aggregation simulations for all other levels) for Greenness, HHeduc, and SES.

Level:	Baseline	One	Two	Three	Four	Five	Six	Seven	Eight	Nine
Greenness										
Coef. mean	(0.320)	0.289	0.287	0.288	0.285	0.285	0.283	0.284	0.283	0.281
Coef. SD		0.016	0.016	0.016	0.017	0.018	0.018	0.018	0.019	0.019
p-value mean	(8.7e-06)	1.2E-04	1.3E-04	1.3E-04	1.6E-04	1.7E-04	2.0E-04	1.9E-04	2.1E-04	2.3E-04
p-value SD		1.2E-04	1.6E-04	1.2E-04	1.7E-04	1.8E-04	2.1E-04	2.0E-04	2.2E-04	2.6E-04
HHeduc										
Coef. mean	(0.038)	0.037	0.038	0.040	0.041	0.042	0.044	0.044	0.045	0.045
Coef. SD		0.003	0.004	0.004	0.005	0.005	0.005	0.005	0.005	0.005
p-value mean	(1.7e-26)	8.5E-09	6.5E-08	3.7E-07	6.0E-06	4.1E-06	3.9E-06	1.0E-05	1.3E-05	2.7E-05
p-value SD		1.1E-07	4.5E-07	2.3E-06	7.2E-05	3.3E-05	2.2E-05	6.7E-05	8.1E-05	2.1E-04
SES										
Coef. mean	(0.175)	0.121	0.101	0.089	0.081	0.072	0.064	0.062	0.055	0.055
Coef. SD		0.021	0.024	0.025	0.030	0.030	0.029	0.032	0.031	0.032
p-value mean	(1.4e-12)	0.003	0.028	0.072	0.137	0.196	0.259	0.299	0.353	0.364
p-value SD		0.007	0.049	0.093	0.161	0.192	0.213	0.240	0.246	0.253

Table 2

Global and local model residual analysis. Mean and standard deviation of normalized Moran's I (i.e., z-scores) from deviance residuals, and mean number of statistically significant High-High (H-H) and Low-Low (L-L) LISA clusters across simulations. Baseline level Moran's I and number of significant clusters from residuals shown for reference (in parentheses).

Level:	Mean Moran's I for Global Residuals	SD of Moran's I for Global Residuals	Mean Moran's I for Local Residuals	SD of Moran's I for Local Residuals	H-H Clusters of Global Residuals	H-H Clusters of Local Residuals	L-L Clusters of Global Residuals	L-L Clusters of Local Residuals
Baseline	(0.041)		(−0.008)		(270)	(173)	(85)	(99)
One	0.091	0.005	0.006	0.004	86.986	55.130	84.8	72.5
Two	0.121	0.007	0.018	0.006	65.342	40.628	62.9	54.0
Three	0.145	0.008	0.029	0.006	50.744	32.546	46.4	37.3
Four	0.165	0.009	0.039	0.008	40.404	27.450	39.2	31.3
Five	0.181	0.010	0.049	0.008	33.904	24.028	33.7	27.1
Six	0.193	0.010	0.055	0.008	28.046	20.570	29.1	24.7
Seven	0.204	0.012	0.062	0.009	23.666	18.450	25.8	22.0
Eight	0.212	0.012	0.068	0.009	20.514	17.142	23.1	19.8
Nine	0.218	0.013	0.074	0.011	18.246	15.984	20.6	17.5