International Journal of Geospatial and Environmental Research

Volume 11 | Number 1 Article 3

May 2024

Urban and Rural BMI Trajectories in Southeastern Ghana: A Space-Time Modeling Perspective on Spatial Autocorrelation

Hsiao-chien Shih
San Diego State University, jxd270700@hotmail.com

Xiaoxiao Wei
Auburn University, xzw0104@auburn.edu

Li An
Auburn University, anli@auburn.edu

John Weeks
San Diego State University, john.weeks@sdsu.edu

Douglas Stow
San Diego State University, stow@sdsu.edu

Recommended Citation

Shih, Hsiao-chien; Wei, Xiaoxiao; An, Li; Weeks, John; and Stow, Douglas (2024) "Urban and Rural BMI Trajectories in Southeastern Ghana: A Space-Time Modeling Perspective on Spatial Autocorrelation," *International Journal of Geospatial and Environmental Research*: Vol. 11: No. 1, Article 3.
Available at: https://dc.uwm.edu/ijger/vol11/iss1/3

Urban and Rural BMI Trajectories in Southeastern Ghana: A Space-Time Modeling Perspective on Spatial Autocorrelation

Abstract

Spatial autocorrelation in model residuals can have a significant impact on the results of spatial or space-time models. This can result in misleading estimates of the influence of different factors, potentially exaggerating or even reversing the perceived effects of these factors. This study also considers the potential implications of the Modifiable Areal Unit Problem (MAUP) in the context of spatial-temporal models. In this case study for southeastern Ghana, we examined whether and how spatial autocorrelation in model residuals might generate bias in regression coefficients when explaining women's body mass index (BMI) across urban and rural areas. Eigenvector spatial filtering, with various settings of influential zones, was systematically tested in a latent trajectory model to detect the impacts of spatial autocorrelation. We found that spatial autocorrelation in model residuals did bias the coefficients of key independent variables such as land cover type, not only affecting their magnitude but also altering their sign or significance. This highlights the risk of significantly misinterpreting the relationships between variables. The bias was effectively mitigated or reduced in urban and rural subsets identified through a data-mining approach, while it persisted in other subsets. This distinction in bias mitigation underscores the necessity of customizing models to suit specific subset attributes. Such systematic testing also enabled our choice of appropriate size of influential zones, within which spatial autocorrelation in data and model residuals was prevalent and thus accounted for biased coefficients. Additionally, we found that BMI trajectories and the associated drivers in urban areas are quite different from those in rural areas, indicating the necessity for differentiating analytical approaches between these areas. This finding therefore justifies the construction of separate BMI models for rural and urban areas. Our methodology demonstrates the importance of managing both temporal variability and spatial autocorrelation simultaneously, improving the model's usefulness in handling other space-time data.

Keywords

spatial autocorrelation; latent trajectory model; eigenvector spatial filtering

Acknowledgements

This research was jointly supported by the National Science Foundation under the Dynamics of Coupled Natural and Human Systems program [DEB-1212183 and BCS-1826839]. Additional funding came from the Interdisciplinary Research in Earth Science (IDS) program of the National Aeronautics and Space Administration (NASA) with NASA Award Number G00009708. San Diego State University and Auburn University also provided financial and research assistance for this study.

1 INTRODUCTION

The increasing prevalence of women's overweight and obesity, a serious problem in Africa, offers a great opportunity to study spatial autocorrelation in model residuals. Before we offer reasons for this statement, we present some background information about obesity challenges in Africa. From 1975 to 2016, prevalence of overweight among African women increased from 13.3% to 34.8% (World Health Organization, 2023). Body mass index (BMI) can be an indicator for people's obesity and vary between urban and rural residents. High BMI may indicate a person is overweight, which can lead to a higher probability of coronary heart disease. Reddy et al. (2002) used BMI to analyze people's health condition in rural and urban areas of north India and found that rural residents had lower BMI compared to urban residents. To effectively mitigate the escalation of obesity prevalence, it is crucial to focus on regions where the issue is most pronounced or where the rate of increase is highest. Developing health policies that prioritize these rapidly expanding areas is essential (Crook et al, 2016).

Given the health concerns related to obesity, especially among West African women, it is crucial to employ advanced analytical methods to comprehend and tackle the concerns. One possible approach to achieve this goal is to implement space-time statistical models to understand how obesity patterns change over space and time and the mechanisms behind such patterns. In the next section, we review the literature related to why and how human socioeconomic and/or behavioral data may be spatially autocorrelated, accounting for spatial autocorrelation in model residuals when such data are used in statistical models. Furthermore, we also show what researchers have done to handle such autocorrelation. Such information helps justify the necessity and the way we have used in this paper related to addressing spatial autocorrelation in model residuals.

1.1 Human Behaviour in Various Living Environments

People living in close geographic proximity tend to display similar habits, lifestyles, and or activity patterns, and such similarity is often reflected in individual- or neighborhood-level data. This situation influences correlation between the attributes of individuals who live in the same or close geographical area(s) at varying degrees, giving rise to spatial autocorrelation (more often used in a geographical context) or neighborhood effects (more often used in a social or economic context). Neglecting these neighborhood effects can lead to biases in modelling results (Zvoleff et al. 2013; An et al. 2016; Sullivan et al. 2017). When considering the neighborhood effects on people's behavioral or health patterns, such effects might vary between rural and urban areas (Glenn and Hill 1977). The behavioral difference can result in disparities in residents' health, demographic features, and socioeconomic status, among others.

1.2 Spatial Autocorrelation in Model Residuals

Spatial data are rapidly being generated and archived over time, yielding spatial-variant, multi-temporal data (space-time data hereafter) that enable the exploration of space-time phenomena and their background mechanisms. Space-time statistical models are extremely useful and becoming increasingly popular for understanding the mechanisms underlying many space-time patterns (An et al. 2015). For example, Rushworth et al. (2014) applied Markov chain Monte Carlo methods to estimate the impact of air pollution on respiratory hospital admissions. Sun et al. (2005)

implemented space-time autoregressive models to accurately estimate housing prices with temporally-dense housing records. For more and different approaches to studying the mechanisms behind many spatially and temporally variant phenomena, refer to An et al. (2016).

Despite advancements in space-time data analysis, the Modifiable Areal Unit Problem (MAUP) remains a significant concern, impacting the interpretability and validity of statistical models. The MAUP stems from the arbitrary aggregation of spatial data, leading to varying statistical outcomes based on the size and shape of spatial units used. This sensitivity to aggregation can lead to misleading conclusions, particularly in spatial regression models when estimating variable relationships (Wong, 2009 ; Manley et al., 2006). Recent studies address the MAUP by utilizing diverse methodological approaches, such as multi-scale analysis to evaluate spatial pattern stability (Maroko et al., 2020) and geographically weighted regression to localize relationships and minimize the impact of spatial aggregation (Fotheringham et al., 2017). These efforts demonstrate the ongoing refinement of strategies to mitigate the effects of MAUP and enhance the robustness of space-time models.

In addition to MAUP, the presence of spatial autocorrelation in such data—and subsequently in model residuals–represents a huge challenge in spatial or space-time data analysis. If not addressed correctly, spatial autocorrelation can undermine the validity of regression outcomes by inflating the significance of coefficients and distorting the standard errors (Griffith, 2000). This inflation and distortion lead to misleading inferences about the relationships between the dependent and independent variables. To improve model performance and associated model coefficient estimation, it is crucially important to reduce or remove autocorrelation in model residuals. Autocorrelation in model residuals results from multiple factors, such as unobserved mechanisms in the model and missing key independent variables (Dormamn et al. 2007). Removing or reducing autocorrelation in residuals can yield a relatively unbiased model. Baltagi (2021) comprehensively summarized multiple approaches to estimating temporal autocorrelation in model residuals. Here, we focus on detecting and reducing spatial autocorrelation in model residuals within the context of space-time statistical models, aiming to correct for these distortions and improve the interpretability and reliability of regression results.

Multiple methods have been proposed to address spatial autocorrelation in model residuals. Cheng et al. (2014) developed dynamic spatial weights matrices and incorporated them in space-time autoregressive integrated moving averages for modeling travel time. However, they did not use global metrics (e.g., global Moran’s I test, Getis’ G test) to examine spatial autocorrelation. Patuelli et al. (2011) modeled German unemployment rates with logistic regression. They found that spatial eigenvectors appropriately handled the spatial autocorrelation in model residuals. This approach primarily addressed the bias in the estimated coefficients, ensuring more accurate representations of the relationships between variables. Gu et al. (2021) applied a negative binomial model with eigenvector spatial filtering (ESF) to model the number of college-graduated workers with variables reflecting economic opportunities and living environment. As a result, spatial autocorrelation in model residuals was greatly reduced and the subsequent model outcomes were more robust in comparison with the ones with distortions typically caused by spatial autocorrelation. In the next section, we explore how spatial autocorrelation may arise in socioeconomic data.

Within the context of examining women's BMI variation over time in Ghana, Crook et al (2016) aimed to remove spatial autocorrelation in model residuals and understand the impacts from sociodemographic and environmental factors. They implemented latent trajectory modeling (LTM) and eigenvector spatial filtering to explain women's BMI with sociodemographic and environmental variables; spatial autocorrelation of model residuals was successfully reduced. They explored Demographic and Health Survey (DHS) and land cover data over space and time and then applied appropriate data preprocessing methods for filling in data gaps which result from repeated cross-sectional surveys. Five latent trajectory models in conjunction with eigenvector spatial filtering were applied to estimate model coefficients and explain women's BMI after removing or minimizing spatial autocorrelation in model residuals. The models explained how women's BMI varied over space and time and the potential impacts from the chosen factors. Land cover composition was found to be an important explanatory environmental factor in the models. A limitation of the Crook et al. (2016) study is that spatial eigenvectors and the size of influential zones were selected in a trial-and-error manner. The optimal number of spatial eigenvectors and the size of influential zones were not determined, and the selection process was somewhat arbitrary.

An influential zone is a geographic area within which an entity influences or is influenced by its spatial neighbors. The size of a spatial influential zone is often determined arbitrarily and without a theoretical or empirical basis (Zvoleff et al. 2013). Spatial adjacency is often applied to polyline and polygon-based data to determine the influential zone (e.g., Patuelli et al. 2011, Rushworth et al. 2014, Cheng et al. 2014). The influential zones remain constant in the Patuelli et al. (2011) and Rushworth et al. (2014) conceptualizations, while the zones in Cheng et al. (2014) vary depending on the traffic speed of line networks. Distance or rank-based approaches are often applied to point-based data. Sun et al. (2005) applied distance-decayed functions to determine whether a condo is within the same building or community of buildings. Aldstadt and Getis (2006) developed a method to determine the size of influential zones based on a multidirectional optimal ecotope-based algorithm, which implements the Getis-Ord local G statistic to identify high and low local centers of phenomena, such as fertility levels (see, for example, Weeks et al, 2010). However, such delicate approaches require complete, longitudinal data over space and time.

In addition, BMI models were assumed to apply similarly to both rural and urban communities. However, Crook et al. (2016) suggested that BMI in urban areas should be modelled separately from rural areas for more accurate BMI models in future studies since factors such as diet and physical activity can vary considerably between urban and rural places. Also, the benefits of treating rural and urban areas differently are observed in other instances (e.g., Glenn and Hill 1977).

1.3 Research Objectives

In this study, the significance of the MAUP issue is recognized by incorporating it into our analytical framework. Building on the recommendation of Crook et al. (2016) study, rural and urban areas are modeled separately. The influence of MAUP on spatial analysis is considered and mitigated by applying eigenvector spatial filtering to filter out and assess the negative impacts of spatially autocorrelated data regarding Ghanaian women's BMI in the corresponding statistical models (detail in Section Data and Methods). Researchers should emphasize the importance of selecting spatial units of analysis that are reflective of underlying processes to avoid potential biases related to the MAUP, an issue not addressed by Crook et al. (2016). Given this context, the goals of this study are to 1) detect the size and impact of influential zones, 2) explore whether and how the eigenvector spatial filtering (ESF) technique can be leveraged to remove or minimize the bias due to neighborhood effects within the relevant influential zones in space-time statistical models, and 3) assess whether new—more plausible—insights regarding the mechanisms behind the BMI space time data can be obtained once the first two goals are achieved.

Compared to Crook et al. (2016), the unique contribution of this study comes from our development of a systematic approach to selecting the optimal number of spatial eigenvectors and the size of influential zones. This approach aids in identifying the appropriate scale of analysis to tackle the MAUP. As pointed out earlier, we separated urban samples from rural samples to investigate the difference in driving forces of women's BMI between urban and rural areas.

2 DATA AND METHODS

2.1 Study Area and Data

The study site is located in southeastern Ghana, West Africa (shown in Figure 1), which coincides with the World Reference System -2 (WRS-2) coordinate in path 193–194, row 55–56. The south coast of Ghana faces the Gulf of Guinea, and its climate, according to Köppen climate classification, belongs to tropical Savanna climate (Aw) that contains wet and dry seasons. Elevation ranges from sea level in the south coast to about 650 m in the north of the study area. The study area contains two major cities, Accra and Kumasi. Agricultural lands surround the two cities, while natural vegetation is the major land cover for the remainder of the study area.

The data for sociodemographic and land cover variables were obtained from two main sources. DHS data were used to represent sociodemographic characteristics. The surveys were conducted at the household level, geo-tagged periodically (typically every five years) and provided at the cluster level. Each cluster was aggregated from multiple individuals and households, so each cluster represents a group of people or households living in the same community. The geographic coordinates of the cluster were randomly shifted (0-2 km for urban areas; 0-5 km for rural areas) from the original location to protect privacy of individuals and households. The DHS data from 1993, 1998, 2003, and 2008 were used to derive sociodemographic variables and track their change over time. At the individual person level, diet and exercise habits are the two main factors that determine whether a person is overweight (Ross et al. 2000). However, DHS survey data did not include any information related to diet or exercise. Therefore, the focus of this study turned to the relationship between sanitation conditions and women's obesity. Sanitation conditions tend to reflect how urban and how wealthy a neighbourhood is. Rural areas are likely not to have a toilet in the house, whereas more affluent urban neighborhoods have a high likelihood of having a flush toilet.

Three variables were extracted and used for modeling women's BMI at the cluster level, including mean women's BMI, % of households with flush toilet (*FlushToilet*), and % of households with no toilet (*NoToilet*). Crook et al. (2016) found significant associations of *FlushToilet* and *NoToilet* with the women's BMI across space and time, so we kept them as covariates in this study. The type of toilet is a function of local infrastructure development. Existence of modern toilet reflects improved environmental and hygienic situations (e.g., lower risk of virus infection), which is one of the factors affecting human weight (Institute of Medicine (US) Subcommittee on Military Weight Management 2004). In addition, each cluster was labeled as either "Rural" or "Urban" in the DHS survey, which was used as an indicator of living environments in the later analysis. Note that BMI was multiplied by 100 to limit the number of digits behind decimal points (*BMI hereafter*).

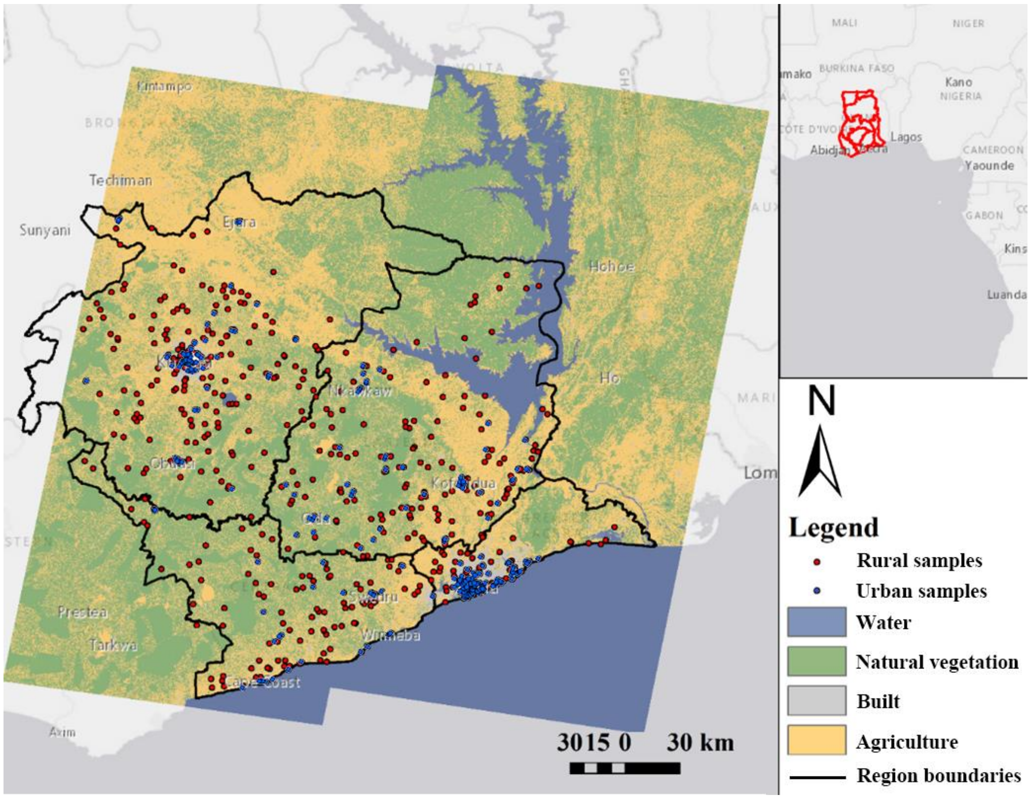

Figure 1. Study area and DHS sample locations. The land cover map was derived from semi-automated classification of a temporal composite (2009-2013) of Landsat imagery by Coulter et al. (2016).

Land cover variables (shown in Figure 1), derived from Landsat satellite imagery, represent the living environments of people surveyed in the DHS samples. Extensive urban expansion and deforestation occurred within the study area, which could affect or be associated with BMI (Crook et al. 2016). Therefore, the land cover variables used in Crook et al. (2016), which accounted for urban expansion and deforestation, were incorporated for modeling BMI in this paper. Because of predominant cloud cover in the study area, multiple dates of Landsat imagery for two epochs (1999-2003 and 2009-2013) were composited to generate land cover maps for the study areas where extensive urban growth and associated land cover change occurred. These two periods were chosen to match DHS data cycles, enabling an integrated analysis of environmental and sociodemographic impacts on BMI. Detailed image processing and land cover derivation are described in Coulter et al. (2016). Each pixel within the study area was classified into one of six subclasses of land cover: water, forest, secondary forest, savanna, agriculture, and built. The five subclasses except water were later aggregated into three more general land cover classes (built, natural vegetation, and agriculture) to represent the distribution and changes in living environments.

2.2 Data Pre-Processing

To extract information for land cover variables, two dates (2000 and 2010) of land cover maps were integrated with the DHS clusters. For each cluster, a 2500 m buffer was created based on its location, where the parameter 2500 m was used to represent a space big enough to include relevant clusters but not too big to include areas that would not influence the focal area. The areal coverage of the built (*Built*) and natural vegetation (*NaturalVeg*) within the buffer were extracted from the two dates of land cover maps, respectively. To estimate the land cover area coverage at the four DHS survey years, linear interpolation (for 2003 and 2008) and extrapolation (for 1993 and 1998) were conducted to generate the land cover data for each date of survey clusters. The data generation and pre-processing are described in detail in Crook et al (2016).

To test whether the spatiotemporal pattern of women's BMI varies over rural and urban areas, subsets of rural and urban samples were generated according to two classification methods (shown in Table 1). First, we labelled each cluster based on DHS classification. The rural and urban clusters were distinguished using a label found in the DHS records. Accordingly, 398 clusters belong to rural samples (DHS rural hereafter), while 382 clusters belong to urban samples (DHS urban hereafter). Second, we adopted another classification method involved gathering rural and urban subsets based on distances to other samples. The spatial distribution data of the samples (shown in Figure 1) indicate that clusters in urban areas are more densely distributed than those in rural areas. However, it is important to address the potential measurement error resulting from the shifting of sampling cluster centroids, which varies between urban and rural areas. Sampling clusters are approximate centroids of areas where groups of households are located, and these centroids can shift over time due to factors such as urban development and rural migration patterns. This shifting can introduce errors in our distance-based classification, as our methodology heavily relies on estimating distances between clusters. To mitigate this, we refined our approach by analyzing the accuracy of cluster designations. We scrutinized how each cluster was determined as urban or rural by examining the criteria we used and comparing them with the actual characteristics of the area. This analysis involved cross-referencing the DHS classification with visual interpretations of the 2000 and 2010 land cover maps.

Furthermore, we acknowledge that using distances between clusters as the primary method for classifying urban and rural areas can be problematic. To address these concerns, we conducted tests and determined various distance thresholds for cluster classification. We set the threshold at less than or equal to 0.5 km for urban clusters and greater than or equal to 3.5 km for rural clusters. This decision was made after carefully considering the spatial distribution data of the samples (Figure 1), which showed a denser distribution of clusters in urban areas.

Table 1. Samples for representing urban and rural subsets in two definitions.

	DHS definition	Sample number	Distance-to-the-nearest neighbor definition	Sample number
Urban	Labeled by DHS survey	382	Distance to the nearest neighborhood cluster smaller or equal to 0.5 km	122
Rural	Labeled by DHS survey	398	Distance to the nearest neighborhood cluster larger or equal to 3.5 km	226

2.3 Latent Trajectory Model

LTM was primarily used to model the temporal trajectory of women's BMI. LTM is a powerful tool for modeling time series data under an assumption that the phenomenon of interest arises from an underlying or latent trajectory over time (Curran et al. 2010). LTM is constructed in a multilevel manner, and some random effects in a LTM can capture the effect of spatial autocorrelation. The shape of the underlying trajectory can be depicted by an intercept (*α*), a slope (*β*), and other optional parameters (e.g., *γ*; Equation 1). *α*, *β*, and *γ* are usually modeled as a function of chosen independent variables (Equations 2~4), which can be either time-variant or time-invariant.

$$BMI_t = \alpha + \beta t + \gamma t^2 + e \qquad (1)$$

The underlying trajectory of women's BMI was modeled as function of an intercept, a slope, and a quadratic term of time (t) (shown in Equation 1). A quadratic term of time was chosen to reflect our observation of some samples having non-linear trajectories of BMI. At *t*=0 (the year 1993), the model reduces to $BMI_t = \alpha + e$, where e represents the error term. This implies that in 1993, the BMI is predicted solely by the intercept α, without additional influence from the linear or quadratic time components. This approach was chosen to simplify the model's initiation, assuming that prior influences up to this point are encapsulated within α and e. For our LTM, α, *β*, and *γ* represent the coefficients of t^0, t^1, and t^2, respectively, and were explained by a set of independent variables.

Our model explains BMI_t using *α*, *β*, and *γ* that represent changes of BMI over time. The values of *α*, *β*, and *γ* already include the influences from earlier times, i.e., $t\text{–}p, \ldots t\text{–}2, t\text{–}1$. For instance, if *γ* is positive, then the trajectory of BMI would have a positive acceleration over time. Put another way, the influences from $t\text{–}p, \ldots t\text{–}2, t\text{–}1$ would affect the values of *α*, *β*, and *γ*, making them to be zero (BMI_t at various times have no temporal influence on one another), positive (positive temporal correlation), or negative (negative temporal correlation). At the same time, our models also allow other variables (i.e., $FlushToilet_t$ and $NoToilet_t$ in Equations 2~4) to affect BMI.

The three parameters *α*, *β*, and *γ* are geographically variant and modeled as a function of several chosen variables in the following equations:

$$\alpha = \alpha_0 + \alpha_1 FlushToilet_t + \alpha_2 NoToilet_t + \varepsilon_0 \tag{2}$$
$$\beta = \beta_0 + \beta_1 FlushToilet_t + \beta_2 NoToilet_t + \varepsilon_1 \tag{3}$$
$$\gamma = \gamma_0 + \gamma_1 FlushToilet_t + \gamma_2 NoToilet_t + \varepsilon_2 \tag{4}$$

where α_0, β_0, and γ_0 are the global intercept, slope, and quadratic coefficient that do not change from cluster to cluster, while the remaining terms (e.g. $\alpha_1 FlushToilet_t + \alpha_2 NoToilet_t$ for α) are determined by the chosen variables that affect coefficients α, β, and γ at the corresponding locations. ε_0, ε_1, and ε_2 are the error terms in these three parameters (α, β, and γ). For model specifications above (Equations 2 ~ 4), the LTM aims to model and explain temporal changes in the dependent variable (*BMI* in our case). If there is no temporal correlation, all *α*, *β*, and *γ* should be zero, implying a stationary situation with no changes over time. However, this contradicts the findings of Crook et al. (2016).

For model A, which is represented in Table 3 as models c0 (urban) and Table 4 as e0 (rural), we used two DHS variables (i.e., *FlushToilet* and *NoToilet*) at date *t* (where $t \in \{0, 1, 2, 3\}$) to estimate the coefficients of *α*, *β*, and *γ* (shown in Equations 2, 3, and 4). To examine whether land cover variables increase model predictive power, we incorporated land cover variables along with the DHS variables in model B, which corresponds to models d1 (urban) in Table 3 and f1 (rural) in Table 4, (shown in Equation 5, 6, and 7). Because the DHS and land cover variables varied over time, BMI at certain points of time were estimated using the DHS and land cover data at the same time (*t*), where $t \in \{0, 1, 2, 3\}$, where 0, 1, 2, and 3 represent 1993, 1998, 2003, and 2008.

$$\alpha = \alpha_0 + \alpha_1 FlushToilet_t + \alpha_2 NoToilet_t + \alpha_3 Built_t + \alpha_4 NaturalVeg_t + \varepsilon_0 \tag{5}$$
$$\beta = \beta_0 + \beta_1 FlushToilet_t + \beta_2 NoToilet_t + \beta_3 Built_t + \beta_4 NaturalVeg_t + \varepsilon_1 \tag{6}$$
$$\gamma = \gamma_0 + \gamma_1 FlushToilet_t + \gamma_2 NoToilet_t + \gamma_3 Built_t + \gamma_4 NaturalVeg_t + \varepsilon_2 \tag{7}$$

2.4 Eigenvector Spatial Filtering

Geographically weighted regression (GWR) is widely used to model many different phenomena with spatial nonstationarity. However, its coefficients tend to demonstrate pronounced multicollinearity and significant positive spatial autocorrelation (Wheeler and Tiefelsdorf, 2005). Griffith (2008) introduced an alternative method to GWR, eigenvector spatial-filter-based local regression, to reduce or remove some of GWR's negative effects. In a particular empirical study, the eigenvector spatial-filter-based local regression technique is suggested superior to GWR due to its enhanced capability to account for autocorrelation in the residuals (Griffith, 2008). The ESF is a spatial filtering method that aims to separate spatially structured components from trend. This distinction is important for addressing the MAUP by taking into account the spatial arrangement of the units of analysis, which helps ensure the reliability of results across varying spatial aggregations. This technique enhances statistical modeling for improved inference and visualization (Griffith et al., 2014).

The ESF approach was incorporated in the LTM to mitigate spatial autocorrelation in model residuals. Spatial filtering methods (Griffith 2000) are instrumental in this context by decomposing key variables in regular multiple regression models to spatial and non-spatial components. The non-spatial components are largely free of spatial autocorrelation, which should be primarily driven by explanatory variables. This method begins with defining a spatial weight matrix, indicating the spatial extent within which samples affect, and are affected by, one another. Spatial autocorrelation can be removed or minimized by incorporating a set of eigenvectors (derived from the spatial weight matrix) as additional independent variables (Griffith 2000, Tiefelsdorf and Griffith 2007, and Chun and Griffith 2011). Each chosen eigenvector represents a spatial pattern of a known or unknown driving force (variable) at a certain scale. In other words, eigenvectors associated with large eigenvalues stand for somewhat large-scale patterns, while those with small eigenvalues stand for small/local patterns (Getis, 2010; Griffith, 2010). Therefore, we do not need to assume the variables (or the subsequent residuals) are spatially autocorrelated in the same way over time.

As we had limited prior knowledge about the size of influential zones in which data are mostly autocorrelated, a data-mining approach was developed to determine the most likely size. Cheng et al (2014) and Aldstadt and Getis (2006) estimated dynamic spatial weight matrices because they had longitudinal data sets, however, our ability to do so is constrained by the temporal incompleteness of the DHS data. Also, the DHS survey density varies over rural and urban areas. The K-nearest neighbor (KNN) method is a conceptually straightforward approach that allows for flexible definition of the influence zone size. It also helps to minimize the negative impacts of spatially autocorrelated data. Thus, the KNN algorithm was used to define the spatial weight matrix (C^k) for all sampled clusters, where k represents the number of nearest neighbors (Aldstadt and Getis 2006). Taking the 2-nearest neighbor definition as an example: for each cluster, 1 was assigned to the nearest two clusters, while 0 to other clusters. The same algorithm was applied to generate spatial weight matrices (C^k) where k represents the tested cluster sizes of 4, 8, 16, 32, and 64, respectively. This process did not group neighborhoods into a single cluster but instead created separate spatial weight matrices for each tested cluster size, treating each as an individual influential zone. A distance-based neighborhood definition was not used because of high variation in density of cluster samples over rural and urban areas. In the original case of the KNN algorithm, the imaginary part in all complex numbers in the eigenvalues were dropped before eigenvector selection, and the top k eigenvectors were selected only based on the real part of associated eigenvalues. To test whether dropping imaginary parts induce substantial differences in regression results, the KNN-spatial weight matrices were forced to be symmetrical to generate the real-number-only eigenvalues and corresponding eigenvectors (forced real number hereafter) for regression analysis.

Given the extensive literature on the selection of appropriate eigenvectors, we employed a data mining approach to systematically determine the optimal number of top k eigenvectors. This approach involved an iterative process where various configurations of eigenvectors were evaluated to ascertain the configuration that minimized spatial autocorrelation in the residuals most effectively, Therefore, we chose the top k eigenvectors (i.e., from E_1 to E_k, where $0 < k \leq 15$ given our moderate sample size; Crook et al. 2016) as independent variables to account for the spatial autocorrelation potentially existing in women's BMI data and independent variables (Griffith 2000). Due to the time-invariant property of spatial eigenvectors, spatial eigenvectors were added to Equations 2~4 and 5~7 for modeling α, β, and γ (An et al. 2016).

3 RESULTS

3.1 Model Fitting

Multiple distance thresholds were tested to classify a certain cluster as urban (if the nearest distance between all pairs of clusters is less than a certain threshold) or rural (the above distance is greater than the threshold). The distance between any two clusters ranged from 0 to 320 km, corresponding to the shortest and longest distances between any two clusters. With this data-driven approach, the urban clusters were derived at the nearest distance less than or equal to 0.5 km, while the rural clusters were derived at the nearest distance greater than or equal to 3.5 km. With these two thresholds, 122 clusters were labeled as urban samples, while 226 clusters were labeled as rural samples. The spatial locations of the 122 and 226 clusters highly reflect the urban and rural environments respectively based on the land cover maps. We excluded 432 clusters with a nearest neighbor distance between 0.5 km and 3.5 km from our data analysis.

To determine the influential zone, the joint effects of two confounding parameters were considered: the number of spatial eigenvectors and the way to define rural and urban sites (Glenn and Hill 1977). To explore whether the land cover variables could substantially increase the explanatory power of the corresponding model, the Akaike information criterion correction (AICc) index was employed to select potential models (Burnham and Anderson 2004).

The trends of AICc were plotted against the size of influential zone and number of spatial eigenvectors (Figure 2). It should be noted that AICc values could not be calculated for certain models due to divergence, resulting in the blank areas in Figure 2(b), (d), (i), and (j). The models with land cover variables have lower AICc over all datasets, which indicates that land cover variables are instrumental in modelling women's BMI.

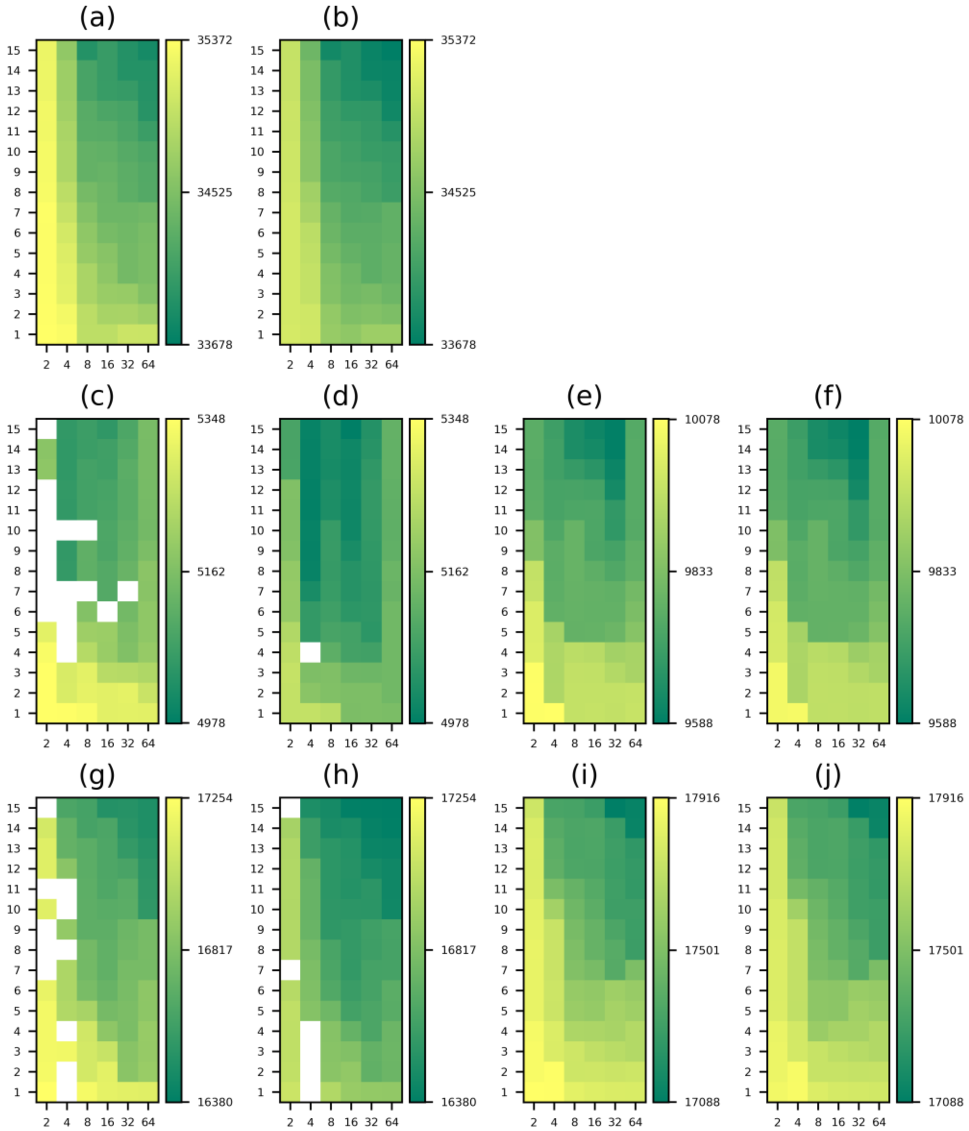

Figure 2. Trends of Akaike information criterion correction (AICc) of BMI models for various settings for influential zones and spatial eigenvectors. (a), (c), I, (g), and (i) are BMI models without land cover variables, while (b), (d), (f), (h), and (j) are BMI models with land cover variables (i.e., built and natural vegetation areas). (a) and (b) were derived from all 780 cluster points; (c) and (d) were derived from NN urban subset that are defined with the nearest distance smaller or equals to 0.5 km(e) and (f) were derived from NN rural subset that are defined with the nearest distance larger or equals to 3.5 km; (g) and (h) are derived from urban subset that are defined by DHS; (i) and (j) are derived from rural subset that are defined by DHS. Vertical axis on each figure represents the number of spatial eigenvectors ranging from 1 to 15. Horizontal axis represents the size of influential zone ranging from 2 to 64, which are shown in scale of log$_2$.

3.2 Spatial Autocorrelation in Model Residuals

The spatial autocorrelation in model residuals was examined with global Moran's I test as shown in Figure 3 and Table 2. The Z-score associated with the global Moran's I dropped with the inclusion of land cover variables over all subsets. Spatial autocorrelation effects were removed or reduced to an acceptable level (i.e., the absolute value of Z-score is less than a threshold such as 1.96) for some models, especially those associated with the NN urban and rural subsets. For the model using the NN urban subset without land cover variables, spatial autocorrelation was removed when the influential zones were 8, 16, 32, and 64 nearest neighbors, along with more than four spatial eigenvectors. Spatial autocorrelation was removed in more combinations of influential zone sizes and spatial eigenvectors for the same subset with land cover variables (i.e., large areas under 1.96; Figure 3(g)). In comparison, spatial autocorrelation was still present for more combinations of influential zone sizes and spatial eigenvectors for the NN rural subset regardless of including the land cover variables or not (see Figure 3(c) and (h)).

In general, spatial autocorrelation was successfully removed, especially for the NN urban subset. However, models with too many spatial eigenvectors may face the 'curse of dimensionality' (Chun and Griffith, 2011), especially when the sample size is limited. In addition, increasing the size of the influential zone may lead to losing the meaning of testing spatial-explicit models. Taking the NN urban subset for example, over half of the total urban samples (64 out of 122) are the spatial neighbors for each sample under the setting of 64-size-influential zone, which implies that spatial autocorrelation at finer scales (i.e., less than 64) is ignored in the corresponding LTM-ESF model.

Table 2. Average Z-Scores associated with global Moran's I test for BMI model residuals over various data subsets. The averaged Z-Scores are derived from all combinations of influential zones and spatial eigenvectors.

	Without land cover variables	With land cover variables
All 780 samples	52.06	49.70
NN urban samples	5.89	4.49
NN rural samples	13.58	13.44
DHS urban samples	22.41	21.93
DHS rural samples	28.72	28.29

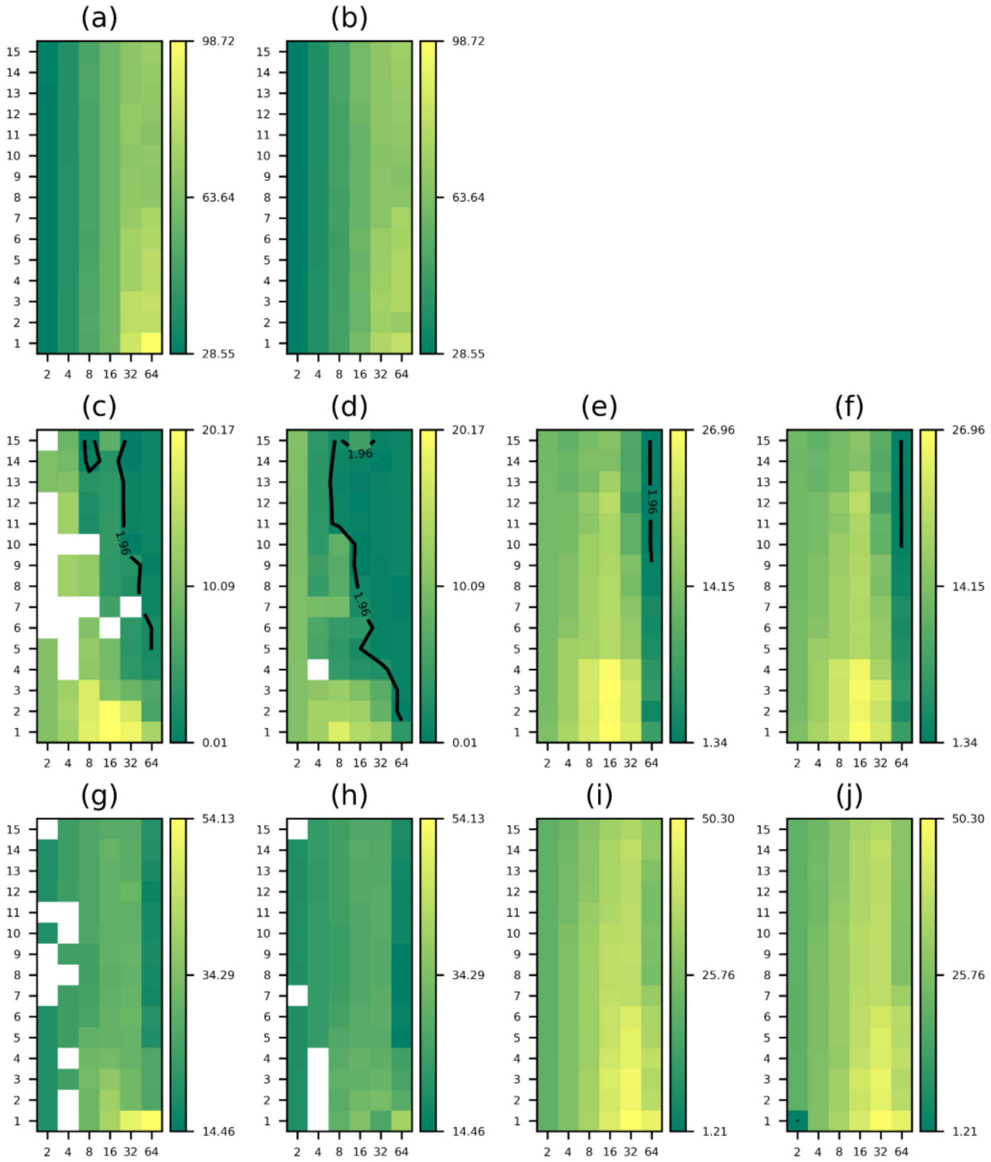

Figure 3. Z-score of BMI model residuals for various settings for influential zone and spatial eigenvectors. Note that (I (c), (e), (g), and (i) are BMI models without land cover variables, while (b), (d), (f), (h), and (j) are BMI models with land cover variables (i.e., built and natural vegetation areas). (a) and (b) were derived from all 780 cluster points; (c) and (d) were derived from the NN urban subset that are defined with the nearest distance smaller or equals 0.5 km; (e) and (f) were derived from the NN rural subset that are defined with the nearest distance larger or equals to 3.5 km; (g) and (h) were derived from urban subset that are defined by DHS; (i) and (j) were derived from rural subset that are defined by DHS. Vertical axis represents the number of spatial eigenvectors ranging from 1 to 15. Horizontal axis represents the size of influential zone ranging from 2 to 64, which are shown in the scale of $\log_2$. Combinations of the size of spatial neighbors and number of spatial eigenvectors that generated a Z-Score less than 1.96 (α less than or equal to .05) are delineated using black contours marked with 1.96 in some of the sub-figures, and sub-figures without the delineation indicate that spatial autocorrelation effects are not removed.

3.3 Optimal Models Without Spatial Autocorrelation

The BMI models without spatial autocorrelation (i.e., Z-score of Moran's I less than 1.96) were identified by analyzing the residuals Moran's I in Figure 3(c) and (d). For the NN urban subset, models with fewest spatial eigenvectors and the smallest influential zones (shown in Table 3) were chosen for model brevity. The BMI models without ESF (i.e., Model c0 and Model d0) are listed for comparison. The coefficient for t^0 (i.e., α_0) is positive and significant for all the four models (Table 3), suggesting that BMI is initially significant and positive. The coefficient β_0 significantly affects all the four BMI models, but γ_0 significantly affects BMI in a declining manner over time. *FlushToilet* and *NoToilet* have significant effects on all models, except for α_2 of *NoToilet*, where α_2 becomes insignificant with spatial eigenvectors. For the NN urban subset with land cover variables (model d0 and d1), the Built land cover variable has a positive impact on the intercept $\alpha 3$, mostly insignificant impact on the slope, but significantly positive impact on the quadratic term. On the other hand, *NaturalVeg* exhibits almost no significant impact in all models. Most spatial eigenvectors significantly influence the BMI, reinforcing the spatially dependent nature of urban BMI trends.

In the NN rural subset, two models were selected and shown in Table 4 along with the models without ESF (i.e., Model e0 and f0). For both models (i.e., Model e1 and f1), the trajectories of BMI show as a convex function of time, indicating that BMI starts at high values (significant, positive α_0), decreases over time (negative, significant β_0), and then increases later (positive, significant γ_0). *FlushToilet* has similar impacts on the intercept, slope, and quadratic term for both the urban (Table 3) and rural (Table 4). *NoToilet* has no significant impact on BMI for the rural subset. Land cover variables consistently have significant effects on the model with *Built* and *NaturalVeg* having almost significant impacts on the intercepts, slope, and quadratic terms of the models. Most spatial eigenvectors have significant effects in the modelling of BMI, and inclusion of the eigenvectors in the model increases the goodness of fit.

4 DISCUSSION

Traditional GIS analytical tools are excellent in handling spatial variability but often struggle with temporal variability. These tools face even greater challenges when it comes to accounting for both spatial autocorrelation and temporal autocorrelation simultaneously (An et al. 2015). The LTM-ESF approach overcomes these limitations by accommodating the complexity of spatial-temporal interactions, offering deeper insights into the underlying mechanisms affecting health outcomes such as BMI. This approach distinguishes our study by elucidating spatial-temporal patterns unique to urban and rural settings, as demonstrated by the divergent trajectories revealed in our analysis. This is an innovative methodological expansion, where traditional spatial analysis and temporal analysis methods are integrated without sacrificing one or the other. The approach is suitable for analyzing certain types of health data, like BMI, which is inherently influenced by both spatial surroundings and temporal changes. However, this methodological expansion is still in its early stages, with many related issues, such as space-time interactions, that require further exploration.

Table 3. BMI models for NN urban subset samples without spatial autocorrelation in residuals

	Model c0 (with spatial autocorrelation)	Model c1 (data-driven urban subset modeled without land cover variables)	Model d0 (with spatial autocorrelation)	Model d1 (data-driven urban subset modeled without land cover variables)
Size of influential zone	0	8	0	8
Number of spatial eigenvectors	0	14	0	11
t^0 (α_0)	2075.52***	2151.75***	1958.52***	2061.73***
FlushToilet (α_1)	2523.56***	1831.54**	2551.15***	2027.75***
NoToilet (α_2)	524.43**	155.53	722.07***	196.25
Built (α_3)			1.1×10-5***	4.981×10^{-6}
NaturalVeg (α_4)			4.615×10^{-6}***	1.300×10^{-5}
t^1 (β_0)	160.00***	137.83***	283.16***	220.38***
FlushToilet (β_1)	-1323.98***	-982.21***	-1204.33***	-1085.12***
NoToilet (β_2)	-615.43***	-298.15*	-842.61***	-366.02**
Built (β_3)			-1×10^{-5}***	-4.640×10^{-6}
NaturalVeg (β_4)			-1.000×10^{-5}	-1.000×10^{-5}
t^2 (γ_0)	-19.54**	-16.04**	-50.57***	-33.92**
FlushToilet (γ_1)	191.15***	144.35***	153.91***	149.98***
NoToilet (γ_2)	151.59***	78.42**	214.07***	93.26**
Built (γ_3)			3.033×10^{-6}***	1.287×10^{-6}***
NaturalVeg (γ_4)			2.686×10^{-6}	2.54×10^{-6}
E_1 (δ_1)		-286.00***		-317.25***
E_2 (δ_2)		-373.73***		-406.40***
E_3 (δ_3)		339.28***		373.20***
E_4 (δ_4)		708.99***		568.02***
E_5 (δ_5)		-74.15***		-106.17***
E_6 (δ_6)		-108.85**		-71.33*
E_7 (δ_7)		195.46**		129.63*
E_8 (δ_8)		295.54**		102.84
E_9 (δ_9)		197.32**		-21.13
E_{10} (δ_{10})		-964.45***		-608.54***
E_{11} (δ_{11})		-554.70***		-371.98***
E_{11} (δ_{12})		-59.17*		
E_{11} (δ_{13})		-48089		
E_{11} (δ_{14})		-48003***		
AICc	5345.3	5076.5***	5258.2	5019.4
Z score of global Moran's I test	>10	0.89	>10	1.43

* means p-value <0.05; ** means -p-value <0.01; *** means p-value <0.0001.

** shaded cells indicate changes in significant level for that correspondent coefficients.

Table 4. BMI models for NN rural subset have no residual spatial autocorrelation.

	Model e0 (with spatial autocorrelation)	Model e1 (data-driven rural subset modeled without land cover variables)	Model f0 (with spatial autocorrelation)	Model f1 (data-driven rural subset modeled with land cover variables)
Size of influential zone	0	2^6	0	2^6
Number of spatial eigenvectors	0	8	0	10
t^0 (α_0)	2135.97***	2161.32***	2089.94***	2090.41***
FlushToilet (α_1)	2258.65***	1733.54***	2204.46***	1884.88***
NoToilet (α_2)	-57.63	-181.72	-44.50	-98.59
Built (α_3)			2.400×10^{-5}**	1.900×10^{-5}*
NaturalVeg (α_4)			3.987×10^{-6}	4.742×10^{-6}*
t^1 (β_0)	1.18	-4.92	46.28	48.10
FlushToilet (β_1)	-828.90***	-587.66**	-757.36**	-634.20**
NoToilet (β_2)	143.02	208.16*	114.96	156.80
Built (β_3)			-2×10^{-5}*	-2.000×10^{-5}*
NaturalVeg (β_4)			-4.170×10^{-6}*	-3.85×10^{-6}
t^2 (γ_0)	12.12**	12.81**	4.97***	3.71
FlushToilet (γ_1)	103.69**	73.95*	77.35	73.87
NoToilet (γ_2)	-39.02	-52.84**	-33.24	-42.61*
Built (γ_3)			3.766×10^{-6}**	3.785×10^{-6}**
NaturalVeg (γ_4)			7.293×10^{-7}***	6.766×10^{-7}***
E_1 (δ_1)		-503.43***		-484.59***
E_2 (δ_2)		86.48**		105.68***
E_3 (δ_3)		-224.05***		-197.88***
E_4 (δ_4)		-71.81**		-86.73**
E_5 (δ_5)		-229.27***		-221.13***
E_6 (δ_6)		98.63**		76.24**
E_7 (δ_7)		41.42		3.60
E_8 (δ_8)		168.39***		451.13***
E_9 (δ_9)				-237.79*
E_{10} (δ_{10})				272.06***
AICc	10077.1	9775.6	10061.1	9754.7
Z score of global Moran's I test	>12	1.91	>12	1.25

* means p-value <0.05; ** means -p-value <0.01; *** means p-value <0.0001.

4.1 Model Difference Between Rural and Urban Samples

By using the selected DHS and land cover variables, the BMI trajectories for rural clusters are substantially different from those of urban clusters. Based on the BMI models derived from the NN urban subset, women's BMI follows a concave function (see Table 3), which indicates that BMI grows gradually, then slows down, and finally decreases in urban areas from 1993 to 2008. Conversely, women's BMI in rural areas has a near zero slope but a positive coefficient for the t^2 term, suggesting an increasing pattern in the long run according to the BMI models from the NN rural subsets (see Table 4). The differences in BMI trajectories indicate that women's BMI will reach (or has reached) a climax for urban areas, which is not observed in rural areas by 2008. Examining these differences is crucial for comprehending the impact of the MAUP, as it reveals spatial patterns that may not be evident when analyzing data in aggregate. The spatial dependence between urban and rural samples was not addressed by Crook et al (2016); our analysis, using separate models for rural and urban samples, reveals significant trajectory differences, suggesting that health outcomes like BMI are not only dependent on individual or community-level factors but also on the larger spatial-temporal context. Our study's methodological approach and findings contribute to a more nuanced understanding of the MAUP, offering a pathway for more localized and effective public health interventions. This insight underscores the need for tailored analytical approaches in spatial studies, considering the diverse characteristics of different geographical segments.

4.2 Challenges in Defining Influential Zones

The detectability of spatial autocorrelation in model residuals can be attributed to several factors. In one instance, this detectability is related to the definition of influential zone. Whether a certain zone really reflects the "neighborhood" within which people influence one another or share some common characteristics is unknown to varying degrees. Thus, the influential zone of each cluster had to be decided based on a data-mining approach (i.e., KNN algorithm) in this study. If the zone reflects the true neighborhood, then we will detect and remove/minimize the spatial autocorrelation through the ESF approach; otherwise, the spatial autocorrelation may still exist even we use the ESF approach.

4.3 Multiscale Spatial Autocorrelation

Another reason why spatial autocorrelation in model residuals is not effectively removed in some instances may stem from multiscale spatial autocorrelation (Overmars et al. 2003), which might reflect the existence of some unobserved independent variables over multiple scales. For example, the residuals of a BMI model are shown in Figure 4, in which the independent variables for the model are DHS and land cover variables along with ESF. A spatial weight matrix was derived from eight-nearest-neighbor influential zone, and the top eight spatial eigenvectors were used in the model. Based on the location of high/low residuals, clusters in the Greater Accra region generally have relatively low residuals, while those around Central and Eastern regions display high residuals. Clusters in the same districts generally have the same level of residuals, which implies that some unobserved independent variables may exist at these two administrative levels. The spatial distribution of residuals clearly shows that rural and urban areas have extensive differences in BMI patterns and the associated mechanisms behind such patterns, which supports our decision to build different models for rural and urban samples. The unobserved independent variables may exert influences on the BMI model, which might arise from differences in regional policies. Thus, BMI models with spatial neighborhood defined at multiple scales (e.g., through putting eigenvectors obtained from different neighborhood definitions into one model) should be explored in future studies.

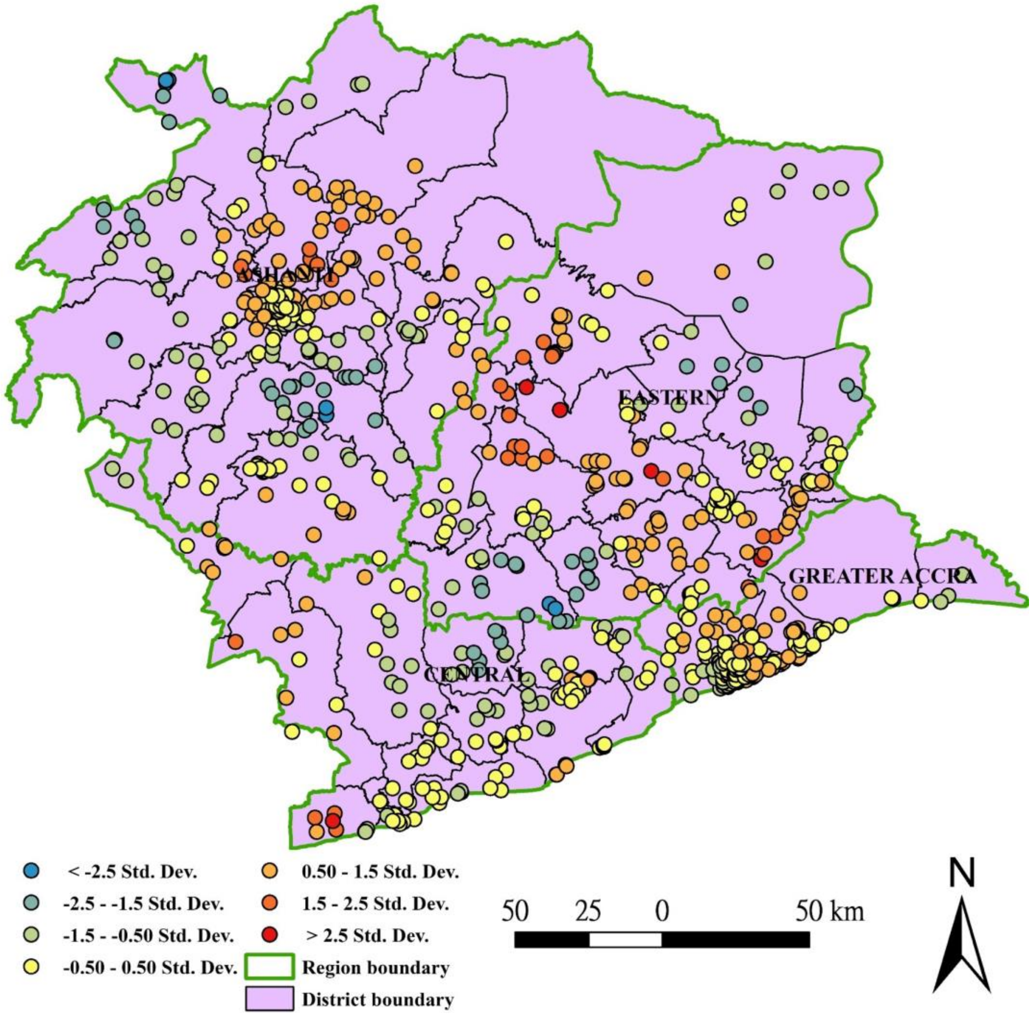

Figure 4. Residuals of a BMI model derived from DHS and land cover variables along with eigenvector spatial filtering based on all 780 DHS samples. The eigenvectors were derived from an eight-nearest-neighbor influential zone, and top eight spatial eigenvectors were used in the BMI model. The residuals are shown in standard deviation fashion.

4.4 Limitations

We suggest that future research focus on the following developments. First, it is recommended to use empirical data if possible. We had quasi-longitudinal data, in which each cluster (location) had interpolated data at four times due to the data limitation. However, such data interpolation, although largely justified in Crook et al. (2016), may still introduce noise or uncertainty. The distinction lies in empirical data being derived from direct observation or experimentation, whereas quasi-longitudinal data, although useful, involves estimation and therefore carries an element of conjecture. This nuance emphasizes the importance of data authenticity in order to minimize analytical distortions. Also, future research could try other methods of selecting spatial filters. In this study, only the top k eigenvectors were adopted although well justified in ESF literature (e.g., Griffith, 2008; Griffith et al., 2014), but it might be worthwhile to test other ways of finding the best eigenvectors, such as the stepwise regression method for spatial eigenvector selection (Griffith 2000, Tiefelsdorf and Griffith 2007). More sophisticated spatial weight matrices (e.g., distance decayed or the k^{th}-order spatial neighborhood definitions) could be tested for the efficacy of removing spatial autocorrelation in model residuals. The eigenvectors are determined by the chosen spatial weight matrix; once it is determined, the eigenvectors become available and do not change over time. Choosing appropriate eigenvectors can help address spatial autocorrelation in future studies.

5 CONCLUSION

This paper presents a data-mining method that aims to detect the effective influential zone size and the number of spatial eigenvectors empirically to support an analysis of spatial-temporal distributions of BMI in Southeastern Ghana. Spatial autocorrelation in model residuals was successfully removed or reduced to an acceptable level for the NN urban and rural subsets of samples. This approach demonstrates the ability to empirically address the MAUP by determining the scale at which spatial processes operate most meaningfully. However, for DHS Urban and Rural subsets, spatial autocorrelation could not be eliminated or reduced. The findings indicate that the model results obtained from the NN urban subsets differ significantly from those of the NN rural subsets, particularly in terms of the sign and significance level of coefficients. Incorporating the unique application of women's BMI in Ghana introduces a new perspective to the research, deviating from conventional analyses. By delineating the size of influential zones and the number of spatial eigenvectors specifically within the context of women's health, our study contributes a novel lens through which to view the interplay of spatial factors in health outcomes. Separating the rural subset from the urban subset was an important step towards uncovering trajectories and the associated mechanisms of BMI changes.

In ideal situations where spatial neighbourhood or influential zone size is known, our LTM-ESF approach will become easier to use—e.g., there is no need to use the data mining method to decide the neighbourhood size, making the modelling practice more straightforward. On the other hand, our LTM-ESF approach can handle instances without prior knowledge about the neighbourhood/zone size, as demonstrated in this paper. Despite the capability of the LTM-ESF approach, we believe that further attention and research efforts—especially in situations with panel data–are necessary for addressing the spatial autocorrelation challenge in space-time statistical models, especially for spatial autocorrelation in model residuals. Such efforts will lead to better capturing the variability in both space and time, revealing the hidden patterns of the phenomena of interest and potential mechanisms.

In conclusion, the LTM-ESF method—in combination with the data mining approach—makes a significant contribution to the field of spatial-temporal analysis, providing a new perspective for examining complex data patterns. Its development signifies progress in methodological approaches, paving an effective way for continued research and refinement to better understand and interpret the intricate dynamics and mechanisms underlying spatial-temporal data.

DATA AND CODES AVAILABILITY STATEMENT

The data and codes that support the findings of this study are available with the identifier(s) at the link https://doi.org/10.6084/m9.figshare.14253122.

ACKNOWLEDGMENTS

This research was jointly supported by the National Science Foundation under the Dynamics of Coupled Natural and Human Systems program [DEB-1212183 and BCS-1826839]. Additional funding came from the Interdisciplinary Research in Earth Science (IDS) program of the National Aeronautics and Space Administration (NASA) with NASA Award Number G00009708. San Diego State University and Auburn University also provided financial and research assistance for this study.

REFERENCES

Aldstadt, J. and Getis, A. (2006) Using amoeba to create a spatial weights matrix and identify spatial clusters. *Geographical Analysis*, 38 (4), 327–343. doi:10.1111/j.1538-4632.2006.00689.x.

An, L., Tsou, M., Spitzberg, B.H., et al. (2016) Latent trajectory models for space-time analysis: an application in deciphering spatial panel data. *Geographical Analysis*, 48 (3), 314–336. 10.1111/gean.12097.

An, L., Tsou, M.-H., Crook, S.E.S., et al. (2015) Space–time analysis: concepts, quantitative methods, and future directions. *Annals of the Association of American Geographers*, 105 (5), 891–914. doi:10.1080/00045608.2015.1064510.

Baltagi, B.H. (2021) *Econometric Analysis of Panel Data*. John Wiley & Sons Ltd: Chichester, England. ISBN: 0-470-01456-3

Burnham, K.P. and Anderson, D.R. (2004) Multimodel inference: understanding AIC and BIC in model selection. *Sociological Methods & Research*, 33 (2), 261–304. doi:10.1177/0049124104268644.

Cheng, T., Wang, J., Haworth, J., et al. (2014) A dynamic spatial weight matrix and localized space–time autoregressive integrated moving average for network modeling. *Geographical Analysis*, 46 (1), 75–97. doi:10.1111/gean.12026.

Chun, Y. and Griffith, D.A. (2011) Modeling network autocorrelation in space–time migration flow data: an eigenvector spatial filtering approach. *Annals of the Association of American Geographers*, 101 (3), 523–536. doi:10.1080/0004

5608.2011.561070.

Coulter, L.L., Stow, D.A., Tsai, Y.-H., et al. (2016) Classification and assessment of land cover and land use change in southern Ghana using dense stacks of Landsat 7 ETM+ imagery. *Remote Sensing of Environment*, 184, 396–409. doi:10.1016/j.rse.2016.07.016.

Crook, S.E.S., An, L., Weeks, J.R., et al. (2016) Latent trajectory modeling of spatiotemporal relationships between land cover and land use, socioeconomics, and obesity in Ghana. *Spatial Demography*, 4 (3), 221–244. doi:10.1007/s40980-016-0024-6.

Curran, P.J., Obeidat, K. and Losardo, D. (2010) Twelve frequently asked questions about growth curve modeling. *Journal of Cognition and Development*, 11 (2), 121–136. doi:10.1080/15248371003699969.

Dormann, C., McPherson, J., Araújo, M., et al. (2007) Methods to account for spatial autocorrelation in the analysis of species distributional data: a review. *Ecography*, 30 (5), 609–628. doi:10.1111/j.2007.0906-7590.05171.x.

Eberhardt, M.S. and Pamuk, E.R. (2004) The importance of place of residence: examining health in rural and nonrural areas. *American Journal of Public Health*, 94 (10), 1682–1686

Fischer, M.M. and Nijkamp, P. (eds.) (2014) *Handbook of Regional Science*. Springer: Berlin, Heidelberg. ISBN: 978-3-642-23429-3.

Fotheringham, A.S., Yang, W. and Kang, W. (2017) Multiscale Geographically Weighted Regression (MGWR). *Annals of the American Association of Geographers*, 107 (6), 1247–1265. doi:10.1080/24694452.2017.1352480.

Getis, A. and Aldstadt, J. (2004) Constructing the spatial weights matrix using a local statistic. *Geographical Analysis*, 36 (2), 90–104. doi:10.1111/j.1538-4632.2004.tb01127.x.

Glenn, N.D. and Hill, L. (1977) Rural-urban differences in attitudes and behavior in the United States. *The ANNALS of the American Academy of Political and Social Science*, 429 (1), 36–50. doi:10.1177/000271627742900105.

Gribov, A. and Krivoruchko, K. (2020) Empirical Bayesian kriging implementation and usage. *Science of The Total Environment*, 722, 137290. doi:10.1016/j.scitotenv.2020.137290.

Griffith, D. (2010) Spatial Filtering. In *Handbook of Applied Spatial Analysis*. pp. 301–318. ISBN: 978-3-642-03646-0.

Griffith, D. and Chun, Y. (2014) Spatial Autocorrelation and Spatial Filtering. In *Handbook of Regional Science*. Springer: Berlin, Heidelberg. pp. 1477–1507. ISBN: 978-3-642-23430-9.

Griffith, D.A. (2000) A linear regression solution to the spatial autocorrelation problem. *Journal of Geographical Systems*, 2 (2), 141–156. doi:10.1007/PL00011451.

Griffith, D.A. (2008) Spatial-filtering-based contributions to a critique of Geographically Weighted Regression (GWR). *Environment and Planning A: Economy and Space*, 40 (11), 2751–2769. doi:10.1068/a38218.

Gu, H., Rowe, F., Liu, Y., et al. (2021) Geography of talent in China during 2000–2015: an eigenvector spatial filtering negative binomial approach. *Chinese Geographical Science*, 31 (2), 297–312. doi:10.1007/s11769-021-1191-y.

Institute of Medicine (US) Subcommittee on Military Weight Management (2004)

Weight Management: State of the Science and Opportunities for Military Programs. National Academies Press (US): Washington DC, US. ISBN-10: 0-309-08996-4.

Krivoruchko, K. (2012) *Empirical Bayesian Kriging: Implemented in ArcGIS Geostatistical Analyst*. https://www.esri.com/news/arcuser/1012/files/ebk.pdf.

Le Gallo, J. (2004) Space-time analysis of GDP disparities among European regions: a markov chains approach. *International Regional Science Review*, 27 (2), 138–163. doi:10.1177/0160017603262402.

Manley, D., Flowerdew, R. and Steel, D. (2006) Scales, levels and processes: studying spatial patterns of British census variables. *Computers, Environment and Urban Systems*, 30 (2), 143–160. doi:10.1016/j.compenvurbsys.2005.08.005.

Maroko, A.R., Nash, D. and Pavilonis, B.T. (2020) COVID-19 and inequity: a comparative spatial analysis of New York City and Chicago hot spots. *Journal of Urban Health*, 97 (4), 461–470. doi:10.1007/s11524-020-00468-0.

Overmars, K.P., De Koning, G.H.J. and Veldkamp, A. (2003) Spatial autocorrelation in multi-scale land use models. *Ecological Modelling*, 164 (2–3), 257–270. doi:10.1016/S0304-3800(03)00070-X.

Patuelli, R., Griffith, D.A., Tiefelsdorf, M., et al. (2011) Spatial filtering and eigenvector stability: space-time models for German unemployment data. *International Regional Science Review*, 34 (2), 253–280. doi:10.1177/0160017610386482.

Reddy, S., Prabhakaran, D., Shah, P., et al. (2002) Differences in body mass index and waist : hip ratios in North Indian rural and urban populations. *Obesity Reviews*, 3 (3), 197–202. doi:10.1046/j.1467-789X.2002.00075.x.

Ross, R. (2000) Reduction in obesity and related comorbid conditions after diet-induced weight loss or exercise-induced weight loss in men: a randomized, controlled trial. *Annals of Internal Medicine*, 133 (2), 92. doi:10.7326/0003-4819-133-2-200007180-00008.

Rushworth, A., Lee, D. and Mitchell, R. (2014) A spatio-temporal model for estimating the long-term effects of air pollution on respiratory hospital admissions in Greater London. *Spatial and Spatio-temporal Epidemiology*, 10, 29–38. doi:10.1016/j.sste.2014.05.001.

Sullivan, A., York, A.M., An, L., et al. (2017) How does perception at multiple levels influence collective action in the commons? the case of Mikania micrantha in Chitwan, Nepal. *Forest Policy and Economics*, 80, 1–10. doi:10.1016/j.forpol.2017.03.001.

Sun, H., Tu, Y. and Yu, S.-M. (2005) A spatio-temporal autoregressive model for multi-unit residential market analysis. *The Journal of Real Estate Finance and Economics*, 31 (2), 155–187. doi:10.1007/s11146-005-1370-0.

Tiefelsdorf, M. and Griffith, D.A. (2007) Semiparametric filtering of spatial autocorrelation: the eigenvector approach. *Environment and Planning A: Economy and Space*, 39 (5), 1193–1221. doi:10.1068/a37378.

Weeks, J.R., Getis, A., Hill, A.G., et al. (2010) Neighborhoods and fertility in Accra, Ghana: an AMOEBA-based approach. *Annals of the Association of American Geographers*, 100 (3), 558–578. doi:10.1080/00045601003791391.

Weeks, J.R., Hill, A.G. and Stoler, J. (2013) *Spatial Inequalities: Health, Poverty, and Place in Accra, Ghana*. Springer: Dordrecht, Netherlands. ISBN: 978-94-007-6731-

7.

Wheeler, D. and Tiefelsdorf, M. (2005) Multicollinearity and correlation among local regression coefficients in geographically weighted regression. *Journal of Geographical Systems*, 7 (2), 161–187. doi:10.1007/s10109-005-0155-6.

Wong, D. (2004) The modifiable areal unit problem (MAUP). In *The SAGE Handbook of Spatial Analysis*. pp. 571–575. ISBN: 978-1-4020-1613-4.

Zvoleff, A., An, L., Stoler, J., et al. (2013) "*What If Neighbors' Neighborhoods Differ? The Influence of Neighborhood Definitions of Health Outcomes in Accra*." In *Spatial Inequalities*. pp. 125-142. ISBN: 978-94-007-6732-4.

APPENDIX

To examine whether complex numbers in eigenvalues diminish the effect of eigenvector spatial filtering, we derived spatial eigenvectors based on a set of distance-based, as well as the "forced-real number" option based spatial weighted matrices and used them as spatial eigenvectors. Such matrices are symmetric, so the resultant spatial eigenvalues are all real numbers. The differences between the original and the "forced-real number" options are minor. Four subsets (NN urban, rural, DHS urban, and rural) were tested for spatial weight matrices derived from distance-based approaches. The n780 dataset was excluded due to the cluster density variation over rural and urban areas. Therefore, we examined whether spatial autocorrelation in model residuals was removed in several additional models with distance-based spatial weighted matrices (weight equals to 1 for samples within distance threshold (dc) from a sample, while 0 otherwise). Distances range from 1 to 6 km were used to determine the influenced zone for NN and DHS urban subsets, while distances from 10 to 60 km were applied to determine the influenced zone for NN and DHS rural subsets. The distance thresholds were determined based on distances to the average nearest neighbor of the rural and urban subsets, respectively. The resultant LTM-ESF models for women's BMI have similar AICc compared to the models with spatial eigenvectors derived from KNN algorithm. However, spatial autocorrelation in model residuals is still present in all models with spatial eigenvector filtering, except for one model in DHS urban subset (1 km of influenced zone and 1 spatial eigenvector). Thus, whether complex numbers are present in the eigenvalues or not does not affect the power of spatial eigenvector filtering.