geographical analysis

Geographical Analysis (2016) **48,** 314–336

Latent trajectory models for space-time analysis: An application in deciphering spatial panel data

Li An[1], Ming-Hsiang Tsou[1], Brian H. Spitzberg[2], Dipak K. Gupta[3], J. Mark Gawron[4]

[1]Department of Geography, San Diego State University, San Diego, CA, USA, [2]School of Communication, San Diego State University, San Diego, CA, USA, [3]Department of Political Science, San Diego State University, San Diego, CA, USA, [4]Department of Linguistics, San Diego State University, San Diego, CA, USA

This article introduces latent trajectory models (LTMs), an approach often employed in social sciences to handle longitudinal data, to the arena of GIScience, particularly space-time analysis. Using the space-time data collected at county level for the whole United States through webpage search on the keyword "climate change," we show that LTMs, when combined with eigenvector filtering of spatial dependence in data, are very useful in unveiling temporal trends hidden in such data: the webpage-data derived popularity measure for climate change has been increasing from December 2011 to March 2013, but the increase rate has been slowing down. In addition, LTMs help reveal potential mechanisms behind observed space-time trajectories through linking the webpage-data derived popularity measure about climate change to a set of socio-demographic covariates. Our analysis shows that controlling for population density, greater drought exposure, higher percent of people who are 16 years old or above, and higher household income are positively predictive of the trajectory slopes. Higher percentages of Republicans and number of hot days in summer are negatively related to the trajectory slopes. Implications of these results are examined, concluding with consideration of the potential utility of LTMs in space-time analysis and more generally in GIScience.

Introduction

Space-time analysis largely refers to detecting, visualizing, or explaining/predicting space-time patterns for certain human or environmental phenomena of interest. As access to large corpora of space-time data has substantially increased, it is evident that space-time analysis has drawn increasing attention. Parallel to this trend, GIScientists face unprecedented challenges and opportunities in "conceptualization, representation, computation, and visualization of space-time data" (Kwan and Neutens 2014, 851). Considerable recent efforts[1] have been devoted to

Correspondence: Li An, Department of Geography, San Diego State University, 5500 Campanile Drive, San Diego CA 92182-4493, USA
e-mail: lan@mail.sdsu.edu

Submitted: June 26, 2015. Revised version accepted: December 04, 2015.

doi: 10.1111/gean.12097

addressing these challenges, including developing and synthesizing space-time conceptual frameworks (An and Crook accepted; Yuan, Nara, and Bothwell 2014; An et al. 2015), eliciting time geography patterns from individual movement and trajectory data (Baer and Butler 2000; Kwan 1998, 2004; Downs, Horner, et al. 2014; Liao, Rasouli, and Timmermans 2014), and developing innovative statistical indices, visualization methods, and/or analytical methods (Rey and Janikas 2006; Chen et al. 2011). For an overview of space-time analysis, including its origin, working definition, historic development, quantitative methods, and strengths and weaknesses, see Yuan, Nara, and Bothwell (2014) and An et al. (2015).

All the above endeavors have undoubtedly contributed to a better understanding of the space-time patterns of interest and the mechanisms behind them. Nonetheless, traditional space-time analysis faces a big challenge largely characterized by its loose coupling of spatial processes and temporal processes of interest. This challenge first lies in the way the space-time phenomenon of interest is represented, and thus the ways in which the corresponding data are collected, organized, and analyzed. Researchers often focus on either the spatial or temporal dimension of the corresponding phenomenon at the expense of the other. Geographers (GIScientists in particular) tend to concentrate on the spatial dimension, which is understandable given their disciplinary emphasis on place and space along with the strength of geographic information systems (GIS) in handling spatial heterogeneity (Peuquet and Duan 1995; An and Brown 2008). However, remedies have been proposed to consider temporal variability in space-time analysis, which, to some degree, have solved the challenge of lacking adequate temporal dimension in space-time analysis (LeSage and Fischer 2008; Elhorst 2012; Downs, Lamb et al. 2014). Evidence can also be found from the development and popularity of the snapshot GIS data model (Armstrong 1988),[2] enabling collection, storage, processing, and analysis of data with both spatial and temporal stamps. Such data largely conform to the so-called spatial panel data or "data containing time series observations of a number of spatial units" (Elhorst 2010, 377).

The limitation in traditional space-time analysis also connects to how spatial panel data are analyzed. Analysis of spatial panel data has in a large sense become nearly interchangeable with space-time analysis. To better explain spatial panel data or make predictions, space-time analysts often use multivariate regression models. The common practice is to consider (1) correlation among cross-time measurements of the same spatial unit, (2) some spatial effect terms such as the spatial lags or spatial errors (Anselin, Le Gallo, and Jayet 2008; Elhorst 2010) or spatial filters (Getis and Griffith 2002), or (3) spatiotemporal autocorrelation such as the geographically and temporally weighted autoregressive (GTWAR) model (Wu, Li, and Huang 2014), which calculates a spatiotemporal distance as a linear combination of both spatial and temporal distances for all the space-time points. Given the resulting spatiotemporal weights matrix that accounts for both spatial and temporal laggings as well as a unique estimation technique, GTWAR is able to fit the space-time data and generate predictions (Wu, Li, and Huang 2014).

Existing analytical methods largely suffer from a lack of robust methods to robustly capture the temporal trend of the phenomena under investigation for each individual spatial unit, particularly as the time span becomes increasingly large. In addition, there is a dire need to better explain (and predict in many instances) such temporal trends. In social science disciplines such as demography and sociology, one approach called latent trajectory modeling (LTM) has demonstrated excellent ability in handling the above challenge of elegantly handling temporal variability in space-time analysis. The objectives of this article are thus threefold: (1) to "borrow" and introduce the LTM approach from social sciences to GIScience, and especially to space-time analysis; (2) to show useful ways to filter out spatial autocorrelation in data before employing the LTM approach; and (3) to demonstrate its power in unveiling the temporal trends hidden in large spatial panel data and the related mechanisms behind such trends.

Under the above three objectives, we introduce the concepts, and demonstrate the usefulness, of LTM using a space time data set collected through several sources. The remainder of this article is organized as follows: The *Methods* section first introduces several key concepts and equations in LTM, justifying why it can be instrumental to space-time analysis. Then, following our data description, the *Results* section reports the outcomes of the data analysis. The *Discussion* section then addresses the implications of these results, and points out several caveats of the work presented in this article. Finally, the *Conclusion* section points out the strengths, weaknesses, and future directions of LTM in GIScience.

Methods

This section introduces LTMs, our space-time data set consisting of cyberspace and realspace variables, and our data analysis strategy.

Latent trajectory models

LTMs are one of the major methods for the analysis of longitudinal data in the social sciences. Also called latent growth models or latent curve models, they are traditionally used to study the (often monotonic) growth of certain study units (e.g., children) in a certain measure (e.g., some test scores) over time (Guo and Hipp 2004; Bollen and Curran 2006, 1–15). The Longitudinal data, characterized by repeated measures for each study unit over time, are assumed to arise from a continuous underlying process or latent trajectory. LTMs aim to use these repeated measures to estimate this latent trajectory that gave rise to these measures: what is the shape of the trajectory, as determined by intercept and slope for a linear trajectory for a linear trajectory? What covariates may help explain the trajectories that vary across study units? Therefore in the LTM approach it is the parameters that characterize these trajectories [e.g., the intercept, slope, and quadratic term in equation (1)], rather than components of these trajectories (i.e., measures at specific times and locations) on which traditional multivariate regression focuses, that we want to explain or predict.

Given the nature of the research question under investigation, the researcher may assume a linear or nonlinear trajectory over time. Among a rich spectrum of linear and nonlinear trajectory models that are readily usable (Bollen and Curran 2006, 88–125), we introduce the quadratic trajectory model that is used later in this article for space-time analysis:

$$y_{it}=\alpha_i+\lambda_t\beta_{1i}+\lambda_t^2\beta_{2i}+\varepsilon_{it} \tag{1}$$

Here the dependent variable y measured for study unit i ($i = 1, 2, \ldots, N$) at time t ($t = 1, 2, \ldots, T$) is denoted as y_{it}, which is modeled as a linear function of an intercept α_i (note: the subscript i indicates this intercept is study unit specific; the same for β_{1i} and β_{2i}), a contribution from time λ_t at the slope β_{1i}, a contribution from the squared time or the quadratic term λ_t^2 at the curvature β_{2i}, and a disturbance ε_{it}. To obtain a linear trajectory model, simply remove the quadratic term $\lambda_t^2\beta_{2i}$ and the model becomes $y_{it}=\alpha_i+\lambda_t\beta_i+\varepsilon_{it}$. The curves below show three exemplar trajectories: linear, quadratic, and exponential (Fig. 1). Note that λ_t represents a linear passage of time, which is often coded as $\lambda_t=0,\ 1,\ 2,\ \ldots,\ T-1$[3] even though there are other

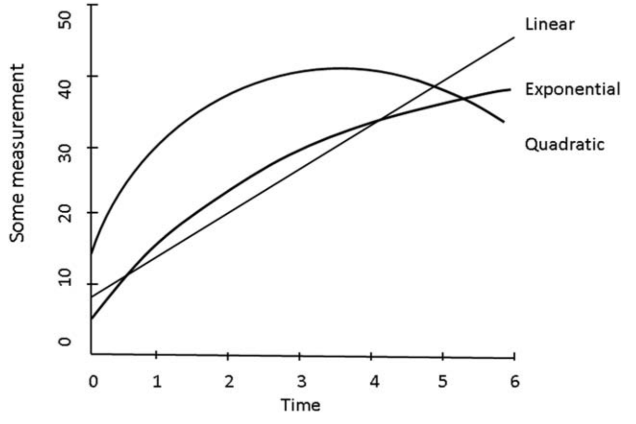

Figure 1. Trajectories of linear, quadratic, and exponential growth.

alternative ways to code time (Bollen and Curran 2006, 113–120). Similarly, λ_t^2 represents the squared value of time, which often takes 0, 1, 4, . . ., $(T-1)^2$ to stand for acceleration (if $\beta_{2i}>0$) or deceleration (if $\beta_{2i}<0$) over time. Equation (1) can also be understood as a way to capture temporal autocorrelation: once the parameters are determined, the y measure at any time can be expressed as a function of the y measure at any other time, the known parameters, the elapsed time, and the time square.

Here we express each trajectory parameter as the sum of an invariant component and a disturbance term, which gives rise to the level 2 equations[4] in regard to equation (1):

$$\alpha_i=\mu_\alpha+\zeta_{\alpha_i} \tag{2}$$

$$\beta_{1i}=\mu_{\beta_1}+\zeta_{\beta_{1i}} \tag{3}$$

$$\beta_{2i}=\mu_{\beta_2}+\zeta_{\beta_{2i}} \tag{4}$$

where μ_α, μ_{β_1}, and μ_{β_2} are the global intercept, slope, and curvature, which are constant across all study units. At the same time, each study unit's trajectory parameters (i.e., the intercept α_i, slope β_{1i}, and curvature β_{2i}) may vary around these three global parameters at a magnitude of three disturbance terms ζ_{α_i}, $\zeta_{\beta_{1i}}$, and $\zeta_{\beta_{2i}}$, respectively. There are a set of assumptions about these disturbance terms about their variance and covariance structures (Bollen 1989, 22, 129), which we skip due to space limitations. The simpler linear trajectory model can be obtained through removing equation (4) and changing β_{1i} to β_i in equation (3).

In parallel with many practices in developing the expansion method (Casetti 1972; Jones, III and Casetti 1992), space-time analysts also tend to include a number of covariates to explain the above trajectory parameters α_i, β_{1i}, and β_{2i} in addition to time and squared-time (or whatever time-related term(s)) as shown above. Assume that two covariates x_1 and x_2 are chosen (for simplicity we do not use the seven variables listed in Table 1). We still use the same level 1[5] trajectory model [i.e., equation (1)], but add $\gamma_{\alpha_1}x_{1i}+\gamma_{\alpha_2}x_{2i}$, $\gamma_{\beta_{1-1}}x_{1i}+\gamma_{\beta_{1-2}}x_{2i}$, and $\gamma_{\beta_{2-1}}x_{1i}+\gamma_{\beta_{2-2}}x_{2i}$ to the right side of equations (1–4), respectively, giving the following three level 2 equations:

$$\alpha_i=\mu_\alpha+\gamma_{\alpha_1}x_{1i}+\gamma_{\alpha_2}x_{2i}+\zeta_{\alpha_i} \tag{5}$$

Table 1. Variables with Data Collected (All Data Are Collected at County Level)

Variable name	Definition	Year	Source
Republican%	Registered Republicans as percent of county total registered voters	2012	Secretary of State (for states with available data) Voter Registration Stats
POP_density	Number of people per squared kilometers*	2010	U.S. Census Bureau; TIGER Line Files of U.S. Census Bureau
POP_Urban	Percent of population that is urban	2000	U.S. Census Bureau
Hot-Days	Annual count of days with daily maximum temperature > 25°C	2006	North American Climate Extremes Monitoring http://www.ncdc.noaa.gov/nacem/index.jsp
Drought-Days	Maximum number of consecutive days without precipitation	2006	Same as above
Age>16%	Percent of population 16 and over	2010	U.S. Census Bureau
Med_HH_Inc	Median household income	2009	American Community Survey

*This measure is derived from total county population and the area of the corresponding county.

$$\beta_{1i}=\mu_{\beta_1}+\gamma_{\beta_{1-1}}x_{1i}+\gamma_{\beta_{1-2}}x_{2i}+\zeta_{\beta_{1i}} \tag{6}$$

$$\beta_{2i}=\mu_{\beta_2}+\gamma_{\beta_{2-1}}x_{1i}+\gamma_{\beta_{2-2}}x_{2i}+\zeta_{\beta_{2i}} \tag{7}$$

Given the increasingly recognized usefulness and popularity of structural equation models (SEM), we use the upper part in Fig. 2 (above the shaded area) to illustrate the quadratic LTM without covariates (represented by equations (1–4)). Readers interested in the details of SEM in this type of applications are referred to Bollen (1989) and Bollen and Curran (2006). The whole diagram (with the lower shaded area) is used to illustrate the model with covariates (represented by equations (1) and (5) through (7)).[6] Repeated measures ($y_1, \ldots, y_5$ shown as rectangles in Fig. 2) are predicted by the trajectory parameters α, β_1, and β_2 (treated as latent variables and shown as circles in Fig. 2) with disturbance terms $\varepsilon_1, \ldots, \varepsilon_5$.

In our model, we assume that data are collected at 5 times to conform to our data set as shown below; in other cases, any number of repeated measures should be fine as long as the research need and model identification requirements Bollen and Curran (2006, 21–25) are satisfied. The numbers near the arrows represent the amount of contribution these trajectory parameters would exert on each y measure. Take y_3 as an example: it receives one unit of contribution from α, two units of contribution from β_1, four units of contribution from β_2, and an unexplained residual term ε_3 (Fig. 2). Put another way, the relationship is $y_3 = \alpha + 2\beta_1 + 4\beta_2 + \varepsilon_3$, which is

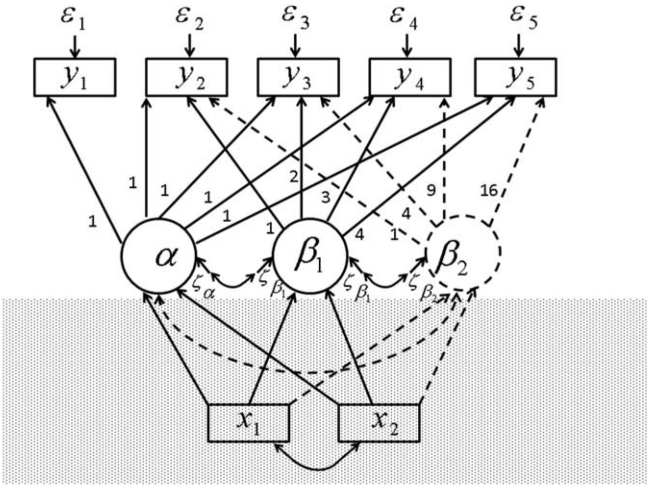

Figure 2. A SEM diagram showing the linear and quadratic LTMs with 5 repeated measures and two covariates (to avoid making the graph too messy, we only show 2 covariates out of the 7 we use). The boxes and arrows in the shaded rectangle are for the LTM(s) with covariates.

equation (1) applied to measures at time 3. Here we also can see the numbers are time steps and squared time steps that have elapsed since the beginning: At time 3, 2 time steps and 4 squared time steps (i.e., 2^2) have passed, which explains the numbers 2 and 4 near the arrows from β_1 and β_2 to y_3 in Fig. 2 as well as those in the equation for y_3.

Similarly, each of the three trajectory parameters or latent variables α, β_1, and β_2 are predicted by covariates x_1 and x_2 (the arrows in the shaded area), and the unexplained parts are expressed as disturbance terms ζ_α, $\zeta_{\beta 1}$, and $\zeta_{\beta 2}$. Therefore, the arrows within the shaded area represent the relationships in equations (5–7). Taking α (or α_i in Fig. 1, where subscript i is used to emphasize α is specific to study unit i) as an example: it is predicted by a global intercept μ_α (not shown in Fig. 2) and three incoming arrows that go from x_1 (at the coefficient of $\gamma_{\alpha 1}$), x_2 (at the coefficient of $\gamma_{\alpha 2}$), and the error term $\zeta_{\alpha.}$ to α. Note the double arrows represent allowance of some covariance structures, for example, between disturbance terms ζ_α and $\zeta_{\beta 1}$, which can be specified in the model.

LTMs, if applied to analyze spatially contiguous data, need to take into account spatial autocorrelation. Put another way, when estimating α, β_1, and β_2 according to equations (5–7) for adjacent or near spatial units, the data at one unit are very likely similar to those at the nearby units. According to Griffith (1988, 2000) and Chun and Griffith (2013), eigenvectors derived from the corresponding binary connectivity matrix can be used as predictors so as to screen out spatial autocorrelation in the data. In practice, this eigenvector spatial filtering (ESF) approach has been demonstrated useful to filter out spatial autocorrelation from relatively small (e.g., dozens or hundreds of spatial units; see Griffith 2002; Helbich and Arsanjani 2015) to relatively big samples (e.g., from 2,352 up to 63,000 spatial units; see Griffith and Fellows 1999; Griffith 2008; Chun and Griffith 2011). Here, we extend the conventional LTM approach and let appropriately chosen eigenvectors enter the model together with the set of covariates (x_1 and x_2 symbolically in Fig. 2). Finally, through the method of maximum likelihood or restricted maximum likelihood, the coefficients and other parameters in equations (5–7) can be estimated.

Once the spatial autocorrelation is addressed in LTMs, the interpolation of model outcomes should focus on equations (5–7), or how covariates may help predict the trajectory parameters α, β_1, and β_2 (represented as straight arrows in the shaded area of Fig. 2). Through affecting these parameters, the covariates may contribute to predicting the dynamical trend of the original measures y_1, y_2, up to y_5. For instance, if a covariate (e.g., x_1) predicts α negatively, β_1 positively, and β_2 negatively, then it suggests that spatial units with high values of the covariate may have low (negative) starting intercepts for y at the beginning (negative coefficient for α) with all other covariates in control, but the y value increases with time (positive coefficient for β_1) in a deceleration manner (negative coefficient for β_2). Similarly, if the coefficients for α, β_1, and β_2 are positive, negative, and positively respectively, then the y measures are high at the beginning, decrease over time, but may finally revert to an increasing trend due to the acceleration term.

Once we have predicted trajectory parameters, we can put them into equation (1), calculate the disturbance ε_{it}, and examine whether these disturbance terms are spatially autocorrelated using existing measures such as Moran's I at each discrete time. As equation (1) is already a function of time (i.e., expressed as λ_t and λ_t^2 in our example), there is no need to examine whether temporal autocorrelation may affect the LTM coefficients and their significance levels as in traditional regression models, where temporally correlated measures are pooled together and put into analysis.

Web-data collection

The LTM approach, when combined with the spatial filtering method, should be ideal for spatial panel data analysis with considerable temporal complexity, including high time frequency, long time span, and/or high spatial or temporal variability in the trajectories of interest. Below we chose to expose this approach to a web-based data set as a demonstration of its usefulness. Our choice of the web-based data set is due to its flexibility in data collection frequency and time span, where realspace data (e.g., population census) do not have such an advantage.

We chose the topic of climate change because few issues loom larger as a threat to human well-being, also because people's interest or concern about climate change may last a long time. Existing literature suggests that concerns or perceptions about climate change are often strongly associated with resource (e.g., water) supply (March, Saurí, and Olcina 2014), regionally relevant activities (Scannell and Gifford 2013), summer temperature or colder winter (Haden et al. 2012), climate variability (Donner and McDaniels 2013), and subnational place attachment or political ideology (Devine-Wright, Price, and Leviston 2015), all of which vary continuously (contingent upon human presence and density) at relatively large scales. Therefore we assume people's concern or care about climate change is spatially continuous and autocorrelated (Tobler 1970) at a relatively large scale. This assumption is also justified by our conceptualization that who you are, what you do, and where you live would affect how much you are influenced by a certain concept or idea (Tsou and Leitner 2013). On the one hand, we are collecting realspace data (i.e., Gallup survey data) over time to show temporal and spatial trends in people's interest and concern about climate change and what factors may affect such trends (An et al. forthcoming).

At the same time, we consider web-based data as a useful data source that may shed important insights into such interest and concerns, and such data as well as the related findings may complement and/or cross-verify the ones from the above realspace data analysis. Under this context, we used Yahoo BOSS search application programming interfaces (APIs) to search webpages relevant to the keyword "climate change" on a weekly basis between 11 November 2011 and 5 March 2013 (Tsou et al. 2013). The Google search engine gives a too limited number of research results per request, which does not satisfy our need to have a relatively large sample so as to produce an interpolated map for the whole continental United States. In addition, the content analysis of our research results for the top 300 websites showed that the Yahoo search returned a higher number of relevant educational websites (see below for why we focused on educational websites), while the Google search missed a set of important educational websites that had been returned by the Yahoo search. Therefore, we chose the Yahoo rather than Google search engine. Based on the returned websites (often in the amount of 700 to 1,000 for each keyword search request) and ranks of these sites given by Yahoo, we calculated a popularity index through reversing the rank value according to the following equation:

$$\text{Popularity measure} = (\text{Total number of web pages} + 1) - \text{rank\#} \tag{8}$$

This popularity measure was calculated for all the returned websites on the following five dates: 11 November 2011 (T1), 4 March 2012 (T2), 1 July 2012 (T3), 4 November 2012 (T4), and 5 March 2013 (T5). This quarterly frequency was chosen based on the aforementioned literature in regard to the influence of climate, weather, resources, and so forth. on people's interest or concern about climate change. All the websites returned from a certain search have an IP address, and relying on the WHOIS database and the geolocation tools developed by (Tsou et al. 2013), we extracted the latitudes and longitudes associated with all IP addresses and the locations of the corresponding web servers. Thus, we mapped the popularity measure of each webpage (web server) for the whole United States.

The issue of dislocation arose, however, in which a website registered at location A may be physically located at (or associated with people who publish their ideas at) location B. For this reason, raw mapping outcomes are not very useful to link to people on the real space. It appears that only around 30% of the registered websites match their physical locations (Lusher 2013). Therefore, we inductively generated a 12-category typology, established its inter-coder reliability, and applied it to classify all the websites into the following categories: blogs, commercial, educational, entertainment and video, forum, governmental websites, informational websites, news, NGO, social media (Twitter, Facebook), special interest, and offline websites (for detail see http://mappingideas.sdsu.edu).

Then because of the relatively high locational accuracy (greater than 74%[7]; Lusher 2013) of educational websites (i.e., website registration sites largely match their physical locations), we mapped the education websites (the URL ends with .edu). These education websites represent all community colleges and 4-year colleges or universities as points on the map of United States along with the popularity measure for each point.

The above points found through the Yahoo search are discrete over space. As mentioned above, people's concern or care about climate change is a large scale, continuous (contingent upon human presence and density), and regionally autocorrelated phenomenon. Therefore, the web-based point data should be interpolated to represent this phenomenon regionally. Among the set of potential interpolation methods such as kriging, kernel density estimation, and various distance-based weighting functions, we chose the kernel density function in GIS because of the many associated benefits, such as allowing one point to represent several observations and fitting a smooth surface around each point with diminishing values until reaching the search radius (zero at the radius; Silverman 1986, 76; ESRI 2013). These properties may better represent the actual situation, in which the concern about climate change may diminish from the corresponding website location till zero (i.e., places with little or none human presence such as national parks and deserts) within a regional scale neighborhood (Scannell and Gifford 2013; Devine-Wright, Price, and Leviston 2015).

To help find out the suitable radius, we did a semi-variogram analysis based on the education website data on March 5, 2013, and found a range that is around 200–500 km (similar results for other four dates). Therefore, we chose a radius of 3 map units as threshold for each kernel (one map unit represents one decimal degree, approximately 80 km in California[8]) and created interpolated, continuous surfaces for the popularity measure at the county level for the whole United States for each of the five chosen dates. This way we may "introduce" large scale spatial autocorrelation in the data (such autocorrelation exists by itself as justified above; we used kernel density function to recover this existence), which might pose a challenge in later regression analysis unless specifically addressed (see section Data visualization and analysis for how we deal with spatial autocorrelation). We chose county as unit of data collection and analysis to allow for linking web-based data to census data (in section Discussion we will discuss this choice).

Counties with high popularity measures, however, do not necessarily imply high interest or concern about climate change. Other factors such as population density, number of websites, and amount of messages posted on websites may confound such measures. To find out the confounding, background "information flow," we randomly selected a set of 168 most frequently used English words in common writings (Tsou et al. 2013). We used these 168 words as keywords, and followed the same procedures as above, and created a background information surface. Subtracting this background surface from the aforementioned five derived surfaces, we obtained five net surfaces that may largely represent the distinctive popularity of climate change from people with, or related to people with, an occupation in education, largely teachers, students, and staff from community colleges, 4-year colleges, and universities. Henceforth, the term *popularity* refers to an interpolated popularity measure [equation (8)] based on detected educational websites. Given the close coupling between people and schools (schools tend to serve, and be located near, people), we interpret the term as the influence of the concept "climate change" (sometimes observed climate change patterns as well) on local people, especially those with or related to an occupation in college education.

Finally, we collected real-space data for seven socioeconomic and demographic variables for all the continental counties in United States from various sources (Table 1). This choice was based on the conceptualization that who you are (demographic features), what you do (income and occupation features), and where you live (geographic, environmental, and climate features) would affect your likelihood of posting ideas or thoughts about an issue on websites (Tsou and Leitner 2013). Based on the unique county ID, these real-space data were linked to the above web-derived popularity measure in ArcGIS.

Data visualization and analysis

We first mapped the above cyberspace data to show how the measured popularity of climate change (among people associated with education occupations) was distributed in the contiguous continental United States over time. Then we picked five exemplar counties, which represent typical temporal trends in the data, to visualize the cross-time variability in the data. Then we built a quadratic trajectory model with and without the above seven covariates (Table 1) to investigate the temporal trajectories following Guo and Hipp (2004).

To address the negative impacts of spatial autocorrelation on regression results (e.g., deflated standard errors and inflated degrees of freedom; Tiefelsdorf and Griffith 2007), we first used a less effective (but helpful in some instances) method for comparative purposes: We randomly chose 25% of all U.S. continental counties (i.e., 791 out of 3,109) and directly used such data in regression analysis. The rationale is that a randomly chosen, small subset of original (often spatially contiguous) data may have spatial autocorrelation at a minimum or substantially reduced (An et al. 2011). On the other hand, eigenvectors can be chosen as predictor variables to filter out such negative impacts of spatial autocorrelation if the whole data set (3,109 counties in our case) is to be used according to Griffith (2000) and Chun and Griffith (2013). One practical way is to choose the top n ($n < 3{,}109$ in our study) eigenvectors[9] such that the Moran's I and the corresponding z score of the regression residuals [ε_{it} in equation (1)] go below a certain threshold value such as 1.96 (Tiefelsdorf and Griffith 2007; Hughes and Haran 2013; Pace, Lesage, and Zhu 2013).

To find out the most appropriate spatial neighborhood definition that largely reflects our choice of a large kernel density radius (section Web-data collection), we tried various orders of contiguity (from 1st up to 20th with all lower-order neighbors included) in GeoDa under the Queen's neighborhood definition, which represent continuously increasing (from the county under consideration) extents of spatial autocorrelation. Under each neighborhood (i.e., a 3,109 $\times$ 3,109 binary matrix of 0s and 1s) definition, we calculated the associated eigenvectors, used a subset of them (up to the top 800 given our large sample size) in combination with the seven covariates (Table 1) as predictor variables, and estimated the corresponding LTM. In each related LTM, the design matrix X includes a column of 1s, the chosen top k (the minimum number k is chosen to lower the spatial autocorrelation to an acceptable level such as z score < 1.96) eigenvectors, and the seven covariates (Table 1).Then under the above LTM, we calculated the corresponding residuals of all regression units (3,109 counties) at the beginning time, that is, 11 November 2011.[10] We then exported these residuals to GeoDa, calculated the Moran's I value and Z score, and evaluated whether the corresponding order of neighborhood is reflective of our previous choice of kernel density radius and also effective to reduce spatial autocorrelation, that is, to lower the z score of the corresponding Moran's I value to acceptable levels (e.g., z score < 1.96).

We first verified the outcome of our spatial neighborhood choice, that is, the correctness of the binary spatial matrices, by examining the 1st, 2nd, and so forth neighborhood of all the 3,109 counties. For several spatial neighborhood definitions (here 3rd, 9th, and 16th),[11] we chose a varying number of eigenvectors (the top 10, 100, 200, 500, and 800, respectively) as predictor variables in conjunction with the abovementioned seven covariates (Table 1). We tested this varying number of eigenvectors to choose an appropriate number of eigenvectors empirically as spatial autocorrelation filters in the context that inclusion of more eigenvectors may decrease the Moran's I value and z score of regression residuals nonlinearly (Tiefelsdorf and Griffith 2007).

Then we implemented the LTM in SAS using the mixed procedure (Guo and Hipp 2004) under two situations: 1) Model 1 (or M#1) in terms of equations (1–4), which includes the chosen eigenvectors, the time (t), and quadratic (t^2) terms; and 2) Model 2 (or M#2) in terms of equations (1), (5), (6), and (7), where the time (t), the quadratic term (t^2), the seven covariates, and the same chosen eigenvectors (as in M#1) were predictor variables. We assumed a constant

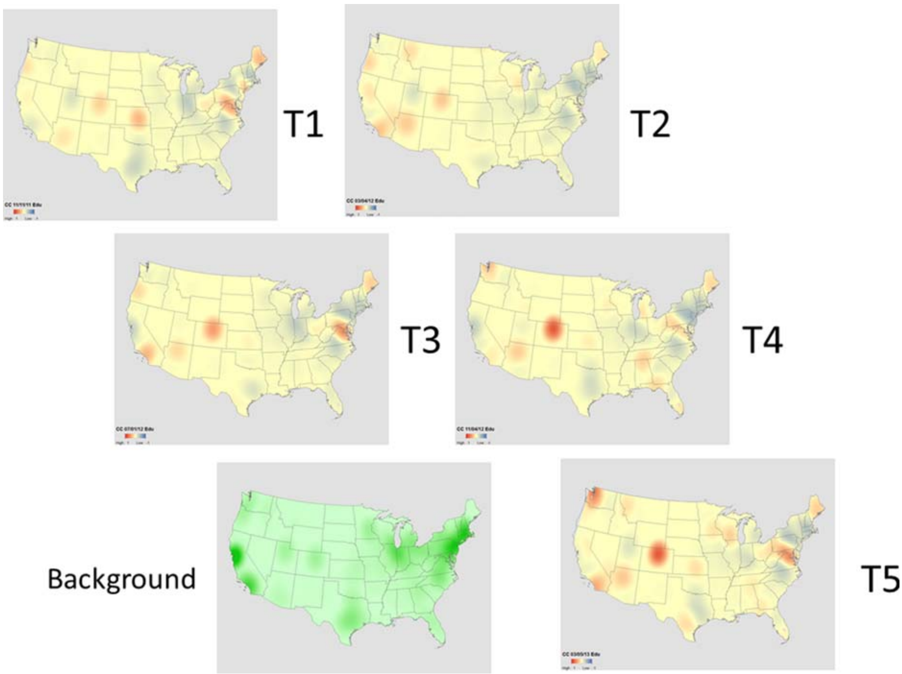

Figure 3. Snapshot maps on the popularity of climate change webpages in education institutions for all continental counties in United States. For what dates T1–T5 represent, see Web data collection section. To make the tone comparable, we have normalized the data before creating the maps. See Appendix II for data normalization detail.

connectivity structure over time (i.e., the correlation in the popularity measure among nearby counties at an earlier time would be carried over to later times), thereby the chosen spatial neighborhood (connectivity) and the derived eigenvectors would be temporally invariant. In the data set with 15,545 (3,109 counties $\times$ 5 times) records, each chosen eigenvector should therefore take the same value over five times for each county.

As the eigenvectors were used as spatial autocorrelation "filters," we did not interpret their coefficients as we did for the rest of the predictor variables. Under both Models 1 and 2 that went through the eigenvector filtering procedure, we assumed a compound symmetry (CS) covariance structure, implying that that all five measurements of the same county over time are correlated in the same (linear or nonlinear) way, and that measures across different counties are not correlated (Guo and Hipp 2004) after the spatial filtering process. For more details about these steps (including the R and SAS code), visit our website at http://complexities.org/LTMs/LTMs.htm.

Results

We first verified our spatial neighborhood choices. Taking San Diego as an example, the 1st, 2nd, 3rd . . .order spatial matrices (output from GeoDa) indicate 3, 7, 15, . . . "neighboring" counties, and the counties thus identified were correct. For instance, Riverside, Orange, and Imperial Counties were identified as the 1st-order neighbors of San Diego; and these three,

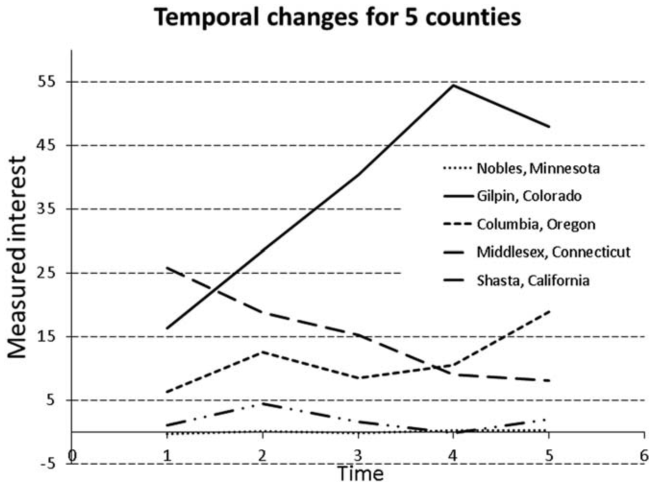

Figure 4. Temporal trajectories of five exemplar counties at five time points between December 2011 and March 2013.

with addition of Yuma (AZ), La Paz (AZ), Los Angeles (CA), and San Bernardino (CA), comprised the seven 2nd-order neighbors, and so forth.[12] Our mapping effort showed that there was considerable spatial variability in the popularity measure. Hot-spots were by and large located in the western part of United States, for example, states of Washington and California, while cold-spots were more in the eastern part. But overall there existed considerable temporal variability (Fig. 3). Note that we did not present the original (before removing the background—see the Appendix about this issue) maps because they are potentially misleading: some hotspots may simply arise from higher background values (e.g., due to more websites, higher population density), and thus are not really hotspots for popularity in climate change. The temporal change in the popularity measure for the five selected counties also suggested substantial variability in the magnitude and change rate of the popularity measure (Fig. 4). When the 3rd-order connectivity was used and the top 200, 500, and 800 eigenvectors were chosen as filters in the LTMs with the seven covariates, the Moran's I (z-value) values were 0.35 (86.91), 0.39 (91.26), and 0.41 (110.61), respectively. Similar results were obtained for the 9th-order neighborhood choices. This may suggest that when spatial connectivity is not big enough to capture the "interpolation-introduced" spatial autocorrelation (a proxy of the intrinsic regional autocorrelation of the phenomenon), relying on a large number of eigenvectors alone is not effective. On the other hand, under a reasonable neighborhood choice, the number of eigenvectors has a nonlinear effect in reducing spatial autocorrelation. For instance, under the16th-order neighborhood choice, the models with top 10 and top 100 eigenvectors gave largely the same regression results in terms of significance level and sign except for some fluctuations in coefficient magnitudes.

After testing over various combinations between the number of eigenvectors (10, 100, 200, 500, and 800) and the order of neighborhood choice (from 1st to 20th), we decided to choose the top 10 (out of 3,109) eigenvectors, under the 16th-order Queen's neighborhood choice (including all the lower order neighbors), as spatial "filters" of spatial autocorrelation in the corresponding LTMs. Doing this way we successfully produced the lowest Moran's I value of 0.0022 (with z varying from 1.95 to 2.47 due to the randomized permutation method for

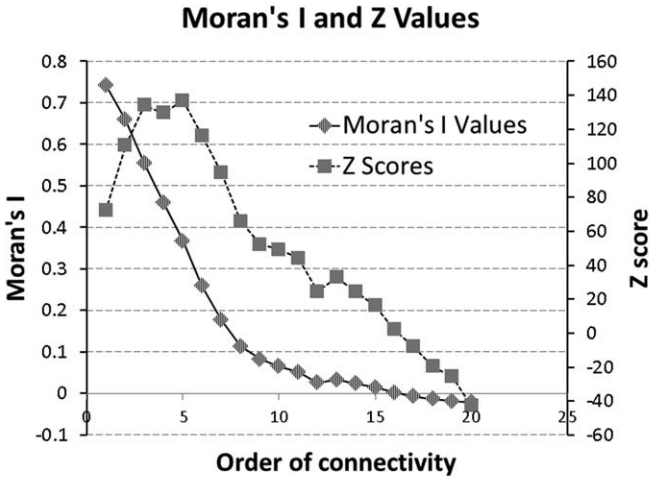

Figure 5. Relationships between Moran's I values, Z scores, and order of connectivity in spatial neighborhood definition.

calculating Z scores in GeoDa; Fig. 5). Under other combinations between the number of eigenvectors and neighborhood choice, we were unable to lower the Moran's I value and z score to this level. The top 10 eigenvectors used in the LTM models (Table 2) were calculated based on this neighborhood definition. To make the results more concise, we only report the regression coefficients of top 4 eigenvectors in Table 2.

Our LTMs revealed some important spatial and temporal trends in the data. Given space limitations, we elaborate on significant variables (i.e., P value < 0.05) only according to the interpretation "rules" set out at the end of section Latent trajectory models. Model 1 showed that all counties started at an average μ_α of 1.6091 [equation (2)], with a linear increase rate (μ_{β_1}) of 1.2713 quarterly (our data were collected at a quarterly interval), but the coefficient for the quadratic term (μ_{β_2}) was negative (-0.1646), implying a deceleration process. This is better understood when we put equations (1–4) back to equation (1): $y_{it} = 1.6091 + \zeta_\alpha \ + \lambda_t$ $(1.2713 + \zeta_{\beta 1}) \ + \lambda_t^2 \ (-0.1646 + \zeta_{\beta 2}) + \varepsilon_{1t}$. The corresponding reduced-form equation, after some algebra transformation, can be rewritten as $y_{it} = 1.6091 + \lambda_t \ (1.2713 - 0.1646 \ \lambda_t)$ + combined error (the eigenvectors are not included to simplify the illustration). This means that as time elapses (i.e., λ_t increases), the term $1.2713 – 0.1646\ \lambda_t$ will decrease till $1.2713 – 0.1646\ \lambda_t = 0$ or $\lambda_t = 7.12$. Once this time threshold is passed, the term $1.2713 – 0.1646\ \lambda_t$ will become negative and thus the measure y_{it} will start to decline. If we graph the above results about Model #1 in conjunction with the average popularity measures from original data at the five time points, these trends could be visualized (Fig. 6).

Model #2 represented the quadratic model with all the time, time-square (quadratic), chosen eigenvectors, and seven covariates (Table 2). The intercept μ_α is -19.0911 ($P = 0.0157$), indicating that if all other variables in equation (5) are set to be zero, counties start at a negative popularity measure at the beginning time. This value, however, is just a regression outcome without much realistic meanings because none of the continental U.S. counties would have all these variables equal zero. Three covariates, population density (0.0042), percent of population 16 years and over (0.1089), and medium household income (0.0002), predicted the intercept α_i positively. These coefficients suggested that the counties with higher values on these three covariates, with control in all other variables, would have higher initial popularity measures at

Table 2. Coefficients and Other Parameters of Three Latent Trajectory Models

Models	#1: Quadratic model (no covariates)	#2: Quadratic model (with covariates)	#3: Quadratic model (with covariates; 25% sample)
Intercept	1.6091 (<0.0001)	−19.0911 (0.0157)	−18.9643 (0.2948)*
t	1.2713 (<0.0001)	−5.4851 (0.2519)	−14.6196 (0.2181)
t^2	−0.1646 (<0.0001)	0.6993 (0.3769)	2.0498 (0.2970)
Predicting intercept			
Republican%		−0.9736 (0.5870)	3.6230 (0.3417)
POP_density		0.0042 (<0.0001)	0.0068 (<0.0001)
POP_Urban		0.0000 (0.9968)	0.0090 (0.6799)
Hot-Days		0.0056 (0.4311)	−0.0064 (0.6027)
Drought-Days		0.0476 (0.2483)	−0.0020 (0.9349)
Age>16%		0.1089 (0.0013)	0.0980 (0.6506)*
Med_HH_Inc		0.0002 (<0.0001)	0.0003 (<0.0001)
Predicting slope			
Republican%		−4.8242 (<0.0001)	−7.2159 (0.0040)
POP_density		−0.0019 (<0.0001)	−0.0041 (0.0005)
POP_Urban		0.0006 (0.9199)	−0.0140 (0.3305)
Hot-Days		−0.0087 (0.0091)	−0.0054 (0.5040)*
Drought-Days		0.0268 (<0.0001)	0.0314 (0.0491)
Age>16%		0.0939 (0.0996)	0.2417 (0.0890)
Med_HH_Inc		0.0000 (0.2160)	−0.0000 (0.3886)
Predicting slope-square			
Republican%		0.9692 (<0.0001)	1.2484 (0.0026)
POP_density		0.0002 (<0.0001)	0.0006 (0.0011)
POP_Urban		0.0000 (0.9712)	0.0016 (0.5131)
Hot-Days		0.0016 (0.0033)	0.0009 (0.4821)*
Drought-Days		−0.0057 (<0.0001)	−0.0051 (0.0516)
Age>16%		−0.0121 (0.1988)	−0.0349 (0.1378)
Med_HH_Inc		−0.0000 (0.0376)	0.0000 (0.4149)*
The first four significant eigenvectors	E1: −56.2045 (<0.0001) E2: −146.87 (<0.0001) E3: −17.3786 (0.0003) E4: −123.64 (<0.0001)	E1: −50.9547 (<0.0001) E2: −86.6602 (<0.0001) E3: −17.6264 (0.2295) E4: −91.8984 (<0.0001)	No eigenvectors used
−2LogL	90156.3	39608.2	10595.9
AIC	90368.3	39682.2	10661.9
BIC	91008.8	39874.1	10789.6

*These variables are shown to have significantly different roles in affecting the intercept, slope, or slope-square due to model specification: Model #3 uses 25% of the whole data records to accommodate spatial autocorrelation without using the eigenvector filtering approach.

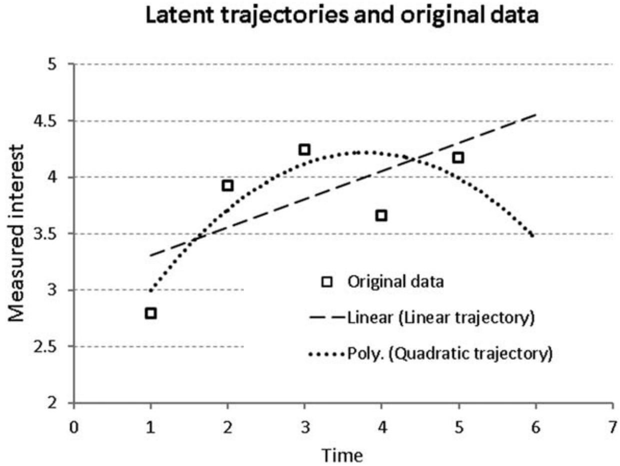

Figure 6. The average trend (popularity of climate change) for all counties in continental United States predicted by the linear model trajectory and by the quadratic model trajectory. The original data at five time points (squares) are also plotted.

the start time [refer to equation (5)). Such information is very useful because important aspects of the space-time data at the beginning of our study time span could be derived: The initial popularity measure for each county [i.e., α_i in Equation (5)] is expressed as a linear combination of a global trend (i.e., μ_α), the covariates multiplied by their corresponding coefficients (i.e., $\gamma_{\alpha_1}x_{1i}+\gamma_{\alpha_2}x_{2i}$[13]), and a random error term (i.e., ζ_{α_i}). Therefore if we know spatial data of the covariates (e.g., maps of the three covariates population density, percent of population 16 years and over, and medium household income), we can easily generate a map of the popularity measure at the initial time.

In predicting the slope or linear term β_{1i} in Equation (6), percent of Republicans (-4.8242), population density (-0.0019), and the number of hot days (-0.0087) were negative predictors, while the number of drought days (0.0268) was a positive predictor. When predicting the quadratic term β_{2i}, percent of Republicans (0.9692), population density (0.0002), and number of hot days (0.0016) positively predicted it, suggesting even though the popularity trajectory may have a decreasing trend in places with high values for these covariates (i.e., they predicted the linear term β_{1i} with negative coefficients of -4.8294, -0.0019, and -0.0087), the decreasing trajectories may turn to an increase curve in places with high values for these three variables. The number of drought days positively predicted the linear term β_{1i} (0.0268), but with a negative coefficient (-0.0057) for the quadratic term β_{2i}, suggesting a county with more drought days would have a more up-shooting trajectory compared to one with fewer drought days, but the up-shooting trajectory may level off and ultimately go down if enough time has elapsed.

Also worthy of mention is the way we handled spatial autocorrelation. Model 3 used only a subset (25%) of all counties in the corresponding LTM without using any of the eigenvector filters following (An et al. 2011). Compared to Model 2, we can see small (moderate to some extent) differences in terms of sign and significance level among the coefficients (Table 2). More specifically, percent of population 16 years and over and the number of hot days in Model 2 were significant predictors for the intercept α_i and slope β_{1i}, respectively, while they became insignificant in Model #3. When predicting the quadratic term, the number of hot days and medium household income changed from significant in Model 2 to insignificant predictors in Model 3.

Discussion

The latent trajectory characteristics revealed by Model #1 agree with our observations. The model without covariates (Model #1; Table 2) predicted a positive intercept and slope, suggesting that the overall popularity was positive at the original time and grew over time. The negative coefficient for time-square, −0.1646, may capture the fact that the popularity measure grows at a decreasing rate over time, and finally the popularity may go down with time. All these agree with the data shown in Fig. 6.

Equally interesting would be the contribution from the covariates. Due to data unavailability, we were not able to collect data for these covariates at (or somewhat prior to) the same five time points. This caveat may not be a big problem if such data do not change substantially over our study time span (we believe this is true for the variables in Table 1), or such temporal changes do not influence the dependent variable substantially. Due to space limitations and our focus on methodology, we only elaborate on the reliability of part of the regression results. Model #2 shows that counties with higher values for percent of Republicans, net the effects of all other covariates, may be associated with a decreasing trend over time as the coefficient of Republican% on β_{1i} negatively (−4.8242) predicted the slope. But this decreasing trend in places with high percent of Republicans would change with time due to the positive coefficient (0.9692) of this variable on the quadratic term. These results suggest that places with higher percentages of Republicans may be less influenced by the concept of climate change at earlier times, but as time elapses, these places might be increasingly influenced. This agrees with the literature in regard to lower percentages of Republicans (compared to Democrats) believing the existence of human-induced climate change (Krosnick, Visser, and Holbrook 1998) and the increasing partisan gap on this belief from 1997 to 2008 (Dunlap and McCright 2008) and from 2001 to 2010 (McCright and Dunlap 2011).

One plausible result is related to the variable number of hot days in summer: Model 2 shows that this variable has an insignificant impact on the intercept α_i (0.0056), but a negative, significant impact on the slope β_{1i} (−0.0087) and a positive, significant impact on the quadratic term β_{2i} (0.0016), indicating that places with varying number of hot days in summer may have largely the same beginning popularity measures, but the trajectory may go down over time in counties with high number of hot days. This downward trend will be reverted due to the positive coefficient for β_{2i} (0.0016), ending up with an upward trajectory when enough time elapses. This phenomenon could be correlated with other factors. For example, counties with higher share in the fossil fuel industry may be thus less influenced by the concept of climate change or the related discussions. Also this may be connected to the fact that these hot places use air conditioning for a long time such that climate change effects (e.g., increasing temperatures) do not cause concerns as large as other places that have fewer hot days and are thus more sensitive to temperature changes.

Overall, the results are very reasonable, and some of them (e.g., those related to the temporal trajectories and predictors) are uniquely obtained using our LTM approach. However, we should point out the caveats of the above regression results.

First the cyberspace data were only from educational websites, and the above results and discussions are only applicable to people related to educational occupation in a strict sense. More effort is needed to establish the link between these educational websites and the general public on the real-space. As previously mentioned, we have been collecting Gallup Poll data to further verify the representativeness of such education website data (An et al. forthcoming). Theories of meme diffusion have contributed to identifying many of the factors specified in this study as relevant moderators of how ideas may diffuse over time and space, including subjective homophily of message content in a social network (e.g., political party), social network centrality (e.g., population density), and geotechnical factors (e.g., weather; Spitzberg 2014). Our model specification should continue to be guided by robust theories and domain knowledge in the future.

Second, the choice of county as unit of data collection and analysis, as mentioned above, was primarily for socioeconomic and demographic data (e.g., U.S. Census data) collection. It should be subject to in-depth test about the modifiable areal unit problem (MAUP) and/or ecological fallacy. Further investigations on other units of analysis (e.g., change the unit to city, township, or census tract) should be carried out in the future. But given the methodological nature of this article as well as its exploratory (data mining) purpose, we hold the value of our data collection and analysis pursuits.

Third and last, we expect that the neighborhood we chose based on the Moran's I and corresponding z values should be largely equal to the extent of spatial autocorrelation specified earlier as the radius (240 km) for the kernel density function, which was empirically determined based on the semivariogram analysis[14] (section Web-data collection). This expectation is confirmed by the large neighborhood chosen above, that is, the 16th order of Queen's connectivity. However given that the sizes of U.S. continental counties are so variable—from 59.13 km^2 of New York County, New York to 51,947.24 km^2 of San Bernardino County, California (https://en.wikipedia.org/wiki/County_statistics_of_the_United_States), further research may choose other potential spatial units that do not vary in area as counties do. At the same time further research should be directed toward examination of the potential negative effects of using such a strongly connected weights matrix, for example, biasing parameter values and significance levels (Smith 2009). Observing that the Moran's I values and the corresponding Z scores largely decrease as the order increases (Fig. 5), some interesting questions may arise (especially) in regard to space time analysis of big data (data often characterized by high volume, velocity, variety, and veracity; http://www-01.ibm.com/software/data/bigdata/, last accessed 16 January 2015): What are the relationships between eigenvector selection and spatial neighborhood definition? Should we still stick to a z score of 1.96 (corresponding to a P value of 0.05) when analyzing big data? More efforts should be invested on such methodological issues in the future.

Conclusions

With the plausibility of our modeling results discussed in section Discussion, we turn to the power of the LTM-ESF approach. First, it is the trajectory of each spatial unit over time, rather than certain measurements within the trajectory, that is of primary interest. When such trajectories are examined in regard to the relevant time-variants and covariates, it is possible to obtain knowledge about the characteristics of, and mechanisms behind, such trajectories. To make the LTM-ESF modeling philosophy easier to understand, we use a three-step procedure to elaborate its robustness and uniqueness: (1) Time-wise regression: regress all values of a certain variable against time given a time series of data at each geographical location, obtain the parameters (e.g., intercept and slope for a linear model of time), and form a trajectory for that location; (2) Trajectory-wise regression: regress these trajectories (or more specifically, the parameters characterizing such trajectories) against a set of selected covariates (which could take varying values over time, but not time per se); (3) Spatial filtering: employ the ESF approach to remove or reduce the biasedness in Step 2 that arises from the spatial autocorrelation among these location-specific trajectories. The above Step 2 for trajectory-wise regression is intellectually similar to earlier practices in developing the expansion method, for example, in relation to expanding the traditional innovation adoption, the population growth, and many other models (Casetti 1972; Jones, III and Casetti 1992).

Given this summary, it is clear that Steps 1 and 2 belong to the traditional LTM, which is powerful enough to handle time series data without considerable spatial autocorrelation. Step 3 comes in when the time series data are also spatially autocorrelated, extending the traditional LTM to geographical analysis with considerable spatial autocorrelation. In handling spatial autocorrelation, it is common that some spatial neighborhood or dependence (e.g., through defining spatial lag, spatial weights matrix, or whatever local neighborhood) is considered in various spatial error, spatial lag, or autoregressive models (Griffith 1988; Getis 1990; Fotheringham, Charlton, and Brundson 2002; Getis and Griffith 2002; Anselin 2003; LeSage and Fischer 2008; Elhorst 2012). The ESF has no exception in this regard. However, the ESF approach is unique in the following aspects. Based on a transformed spatial weights matrix, mutually orthogonal and uncorrelated eigenvectors can be derived to furnish uncorrelated map patterns, ranging from the largest possible positive to the largest possible negative spatial autocorrelation levels (Griffith 2000; Tiefelsdorf and Griffith 2007; Chun and Griffith 2011). Then through applying a spatial filter (i.e., a subset of eigenvectors obtained through stepwise regression or selected from the ones with the highest eigenvalues) to the regression model as independent variables, the spatial components in the related variable can be accounted for, resulting in unbiased parameter estimates (Tiefelsdorf and Griffith 2007). The ESF approach to handling spatial autocorrelation has many benefits, including applicability to all data types (e.g., binary data, categorical data) and flexibility in choosing various form spatial neighborhood definitions (e.g., the n nearest neighbors, within a certain distance, the ones we used in this article). For more detail about this topic see An et al. (2015).

Additionally, this LTM-ESF approach has superior capacity to handle temporal variability compared to most other modeling approaches. For instance, if data at all times are used in traditional ordinary-least-squares (OLS) regression, the assumption of interdependence among observations will be violated as multiple measurements of the same spatial units over time are very likely correlated. Alternatively, we may choose data at one time or use the average number (over time) for each spatial unit in regression, which loses the time dimension and thus, in essence, is not space-time analysis. In handling temporal autocorrelation, the LTM-ESF approach is unlike many geographical approaches (e.g., GTWAR as reviewed in Section Introduction) that use time lags or temporal weights such that data at nearer time points contribute more to predicting the value at the time of interest. Instead, data points at all times contribute to constructing a trajectory for each location of interest. This type of way of handling temporal relationships in data should be more powerful when the time span is relatively long, the temporal resolution is very fine, and/or outliers, errors, or variability in measurements are prevalent at various or some time points. This not only saves computational time (e.g., no need to construct and compute so many local temporally weighted matrices) when dealing with big data, but also reduces the sensitivity and vulnerability of regression results to measurement errors (the whole trajectories, not data at these time points, are subject to analysis).

Therefore, the LTM-ESF approach empowers modelers to simultaneously take into account both spatial (the ESF part) and temporal (the LTM part) autocorrelations, making space-time analysis more capable of handling empirical space time data, especially big space time data. As shown in Section Results, the coefficients predicting the intercept α of the related LTM model can be used to derive important aspects of the corresponding space-time processes at the beginning time. This merit may suggest that LTM could be a powerful tool for relatively tight coupling of both spatial processes and temporal processes of interest.

The models in this article can be improved in several aspects. First, our LTMs were programmed in SAS, which does not provide several overall fit indices that are available in other alternative software packages such as LISREL and MPLUS (Guo and Hipp 2004; Bollen and Curran 2006, 44–54). Further testing the above models in these software packages may help improve confidence in the model outcomes. Second, in regard to different ways of handling spatial autocorrelation, we would lean toward Model #2 given our possibly more effective handling of spatial autocorrelation (Moran's $I = 0.0022$ with a z score near 1.96) using the eigenvector approach (computationally intensive though) than the subsampling approach used in Model #1. On the other hand, we believe that the subsampling approach is still useful in reducing spatial autocorrelation given its ease of use and its capacity to mostly reveal the effects of predictable variables on the popularity trajectories. This method may find more applications in analysis of big data, in which dropping some records during the subsampling process is not only affordable but also sometimes helpful (e.g., in computationally intensive applications) due to abundance of data. Third, this article demonstrates the usefulness of LTMs only using the linear and quadratic forms, but there are other functional trajectories (piecewise, exponential, cosine, etc.) or combinations of such forms that can be applied in other space-time analysis applications.

LTMs may find applications in many areas within and outside the GIScience domain. Examples include multitime analysis of some vegetation indices (e.g., NDVI after cross-time registration) based on satellite imagery, space-time analysis of Twitter data (e.g., for detecting diseases; Nagel et al. 2013), and regional space-time inequality analysis (e.g., for economic development; Rey 2001; Rey and Janikas 2006). The rule of thumb is that, as long as spatial panel data are available with autocorrelation existing over space and time,[15] these data can be exposed to this powerful approach. It is our hope that this article may remind GIScienst (or scientists in general) of the usefulness of the LTM approach in deciphering spatial panel data. We believe that more efforts should be invested to exploring its strengths, weaknesses, and application domains in many geographic or interdisciplinary studies.

Appendix : Data normalization

For a certain time, we divide each county's measure (based on classified data shown in section Web-data collection) by the maximum measure among the whole 3,109 counties. This gives rise to ratio 1 ($R1$), Then we divide each county's noise measure (also based on classified data shown in section Web-data collection) by the maximum noise measure among the whole 3,109 counties. This gives rise to ratio 2 ($R2$). The values in Fig. 3 are the differences between these two ratios, or $R1 - R2$. But the data used in analysis are not subject to this normalization.

Notes

1 The number of publications in relation to space-time analysis, according to an online search from web of knowledge, has experienced an exponential increase since around 1990–1993 (An et al. 2015).
2 See related literature for strengths and weaknesses of this data model (Peuquet 1994; An and Brown 2008) and other alternative data models (Langran and Chrisman 1988; Worboys 1992; Yuan 1999).
3 A convention in LTMs is to code the beginning time as 0, and treat the first measure as the initial value or intercept, and the following measures as a linear combination of this initial value and changes brought in by the amount of time (and squared-time) that has passed since the beginning (Bollen and Curran 2006, 90–93).
4 If substituting Equations (2)–(4) for the corresponding terms in equations (1), we arrive at the so-called combined or reduced-form model. For simplicity we skip it, but the reader should be able to derive it using simple algebra.
5 LTMs share much overlap (even equivalence) with multilevel models in many instances (Guo and Hipp 2004; Preacher et al. 2008). Our SAS code is programmed in the *mixed* procedure that is also used for multilevel modeling.
6 If the dashed circle and arrows in Fig. 2 are removed, the diagram becomes the linear LTM as described above.
7 This accuracy was calculated at the city level (Lusher 2013). If aggregated to the county level (the spatial unit of this article), the accuracy should reach around 85% according to the related literature (e.g., Shavitt and Zilberman 2010) and our team's similar search on other keywords (Tsou et al. 2013).
8 This choice also hinges upon a communication theory about idea diffusion (Spitzberg 2014) as well as our observations that different phenomena are spatially autocorrelated at different spatial extents and levels (An et al. forthcoming).
9 We did not perform stepwise regression to choose eigenvectors as regressors as Tiefelsdorf and Griffith (2007) did for two reasons: (1) choosing the top n (n could be 10, 100, 200, etc. in our case) eigenvectors could also be appropriate, especially in situations with a large number of spatial units according to our personal communication with Griffith and Chun (2014); (2) our data analysis later also showed that the inclusion of the top 10 eigenvectors would reduce the spatial autocorrelation to an acceptable level.
10 The residuals associated with the other four times, when exposed to the Moran's I test, show very similar patterns as those at the beginning time.
11 These orders were chosen based on some preliminary observations that a neighborhood defined at a low order alone was not likely to effectively reduce spatial autocorrelation to an acceptable level (e.g., $z < 1.96$) no matter how many eigenvectors were included. This may arise from the large spatial autocorrelation size defined earlier in the radius of the kernel density function (section Web-data collection).
12 As mentioned earlier, our neighborhood definition is inclusive of all lower-order neighbors. For instance, when creating the eigenvectors for the 2nd-order spatial neighborhood, all the 7 neighbors (three 1st- and four 2nd-order neighbors) were used.
13 As mentioned earlier in section Latent trajectory models, this is a symbolic representation of all (not necessarily 2) that can be included.
14 There is a growing amount of literature about extending traditional geostatistical methods developed for point data (e.g., semivariogram) to applications in social sciences and medical sciences, which has enabled these methods to handle aggregate data on regular or irregular area units such as census units and counties (Goovaerts 1997, 2008; Chiles and Delfiner 1999).
15 Latent trajectory models allow some temporal complications in data collection and analysis, such as unequal spacing over time (e.g., data collected at years 1, 2, 5, 6, and 9) and unbalanced data (some units do not have data at some time points (Guo and Hipp 2004).

References

An, L., and D. G. Brown. (2008). "Survival Analysis in Land Change Science: Integrating with GIScience to Address Temporal Complexities." *Annals of the Association of American Geographers* 98(2), 323–44.

An, L., D. G. Brown, J. I. Nassauer, and B. Low. (2011). "Variations in Development of Exurban Residential Landscapes: Timing, Location, and Driving Forces." *Journal of Land Use Science* 6(1), 13–32.

An, L., and S. Crook. accepted. "Spatiotemporal Analysis." *The International Encyclopedia of Geography: People, the Earth, Environment, and Technology*.

An, L., M.-H. Tsou, S. Crook, Y. Chun, B. Spitzberg, J. M. Gawron, and D. Gupta. (2015). "Space-Time Analysis: Concepts, Quantitative Methods, and Future Directions." *Annals of Association of American Geographers* 105(5), 891–914.

An, L., M.-H. Tsou, B. Spitzberg, D. K. Gupta, S. Wandersee, and J. M. Gawron. (forthcoming). "Climate Change: Who Cares?

Anselin, L. (2003). "Spatial Externalities, Spatial Multipliers, and Spatial Econometrics." *International Regional Science Review* 26(2), 153–66.

Anselin, L., J. Le Gallo, and H. Jayet. (2008). "Spatial Panel Econometrics." In *The Econometrics of Panel Data: Fundamentals and Recent Developments in Theory and Practice, Advanced Studies in Theoretical and Applied Econometrics*, 901–69, edited by L. Mátyás and P. Sevestre. Berlin, Germany: Springer-Verlag.

Armstrong, M. P. (1988). "Temporality in Spatial Databases." In *GIS/LIS 88 Proceedings: Accessing the World*, 880–89. Falls Church, VA: American Society for Photogrammetry and Remote Sensing.

Baer, L. D., and D. R. Butler. (2000). "Space-Time Modeling of Grizzly Bears." *The Geographic Review* 90(2), 206–21.

Bollen, K. A. (1989). *Structural Equations with Latent Variables*. New York: Wiley.

Bollen, K. A., and P. J. Curran. (2006). *Latent Curve Models: A Structural Equation Perspective*. New Jersey: Wiley.

Casetti, E. (1972). "Generate Models by the Expansion Method: Applications to Geographical Research." *Geographical Analysis* 4(1), 81–91.

Chen, J., S.-L. Shaw, H. Yu, F. Lu, Y. Chai, and Q. Jia. (2011). "Exploratory Data Analysis of Activity Diary Data: A Space–Time GIS Approach." *Journal of Transport Geography* 19(3), 394–404.

Chiles, J.-P., and P. Delfiner. (1999). *Geostatistics: Modeling Spatial Uncertainty*. New York: Wiley.

Chun, Y., and D. A. Griffith. (2011). "Modeling Network Autocorrelation in Space–Time Migration Flow Data: An Eigenvector Spatial Filtering Approach." *Annals of the Association of American Geographers* 101(3), 523–36.

Chun, Y., and D. A. Griffith. (2013). *Spatial Statistics and Geostatistics: Theory and Applications for Geographic Information Science and Technology*. Los Angeles: SAGE Publications Ltd.

Devine-Wright, P., J. Price, and Z. Leviston. (2015). "My Country or My Planet? Exploring the Influence of Multiple Place Attachments and Ideological Beliefs Upon Climate Change Attitudes and Opinions." *Global Environmental Change* 30, 68–79.

Donner, S. D., and J. McDaniels. (2013). "The Influence of National Temperature Fluctuations on Opinions About Climate Change in the U.S. Since 1990." *Climatic Change* 118, 537–50.

Downs, J. A., D. Lamb, G. Hyzer, R. Loraamm, Z. J. Smith, and B. M. O'Neal. (2014). "Quantifying Spatio-Temporal Interactions of Animals Using Probabilistic Space–Time Prisms." *Applied Geography* 55, 1–8.

Downs, J. A., M. W. Horner, G. Hyzer, D. Lamb, and R. Loraamm. (2014). "Voxel-Based Probabilistic Space-Time Prisms for Analysing Animal Movements and Habitat Use." *International Journal of Geographical Information Science* 28(5), 875–90.

Dunlap, R. E., and A. M. McCright. (2008). "A Widening Gap: Republican and Democratic Views on Climate Change." *Environment* 50(5), 26–35.

Elhorst, J. P. (2010). "Spatial Panel Data Models." In *Handbook of Applied Spatial Analysis: Software Tools, Methods and Applications*, 377–407, edited by M. M. Fischer and A. Getis. Berlin: Springer-Verlag.

Elhorst, J. P. (2012). "Dynamic Spatial Panels: Models, Methods, and Inferences." *Journal of Geographical Systems* 14(1), 5–28.

ESRI. (2013). *ArcGIS*. Redlands, CA: Environmental Systems Research Institute. Available online at http://www.esri.com/.

Fotheringham, A. S., M. Charlton, and C. Brundson. (2002). *Geographically Weighted Regression: The Analysis of Spatially Varying Relationships*. Chichester, UK: Wiley.

Getis, A. (1990). "Screening for Spatial Dependence in Regression Analysis." *Papers of the Regional Science Association* 69(1), 69–81.

Getis, A., and D. A. Griffith. (2002). "Comparative Spatial Filtering in Regression Analysis." *Geographical Analysis* 34(2):130–40.

Goovaerts, P. (1997). *Geostatistics for Natural Resources Evaluation*. New York: Oxford University Press.

Goovaerts, P. (2008). "Kriging and Semivariogram Deconvolution in the Presence of Irregular Geographical Units." *Mathematical Geology* 40(1), 101–28.

Griffith, D. A. (1988). "Estimating Spatial Autoregressive Model Parameters with Commercial Statistical Packages." *Geographical Analysis* 20(2), 176–86.

Griffith, D. A. (2000). "A Linear Regression Solution to the Spatial Autocorrelation Problem." *Journal of Geographical Systems* 2(2), 141–56.

Griffith, D. A. (2002). "A Spatial Filtering Specification for the Auto-Poisson Model." *Statistics & Probability Letters* 58, 245–51.

Griffith, D. A. (2008). "Spatial-Filtering-Based Contributions to a Critique of Geographically Weighted Regression (GWR)." *Environment and Planning A* 40, 2751–69.

Griffith, D. A., and P. Fellows. (1999). "Pixels and Eigenvectors: Classification of LANDSAT TM Imagery Using Spectral and Locational Information." In *Spatial Accuracy Assessment: Land information Uncertainty in Natural Resources*, 309–17, edited by K. Lowell and A. Jaton. Ann Arbor, MI: Ann Arbor Press.

Guo, G., and J. Hipp. (2004). "Longitudinal Analysis for Continuous Outcomes." In *The Handbook of Data Analysis*, 347–68. edited by M. Hardy and A. Bryman. Los Angeles: SAGE Publications.

Haden, V. R., M. T. Niles, M. Lubell, J. Perlman, and L. E. Jackson. (2012). "Global and Local Concerns: What Attitudes and Beliefs Motivate Farmers to Mitigate and Adapt to Climate Change?" *PLoS One* 7(12), e52882.

Helbich, M., and J. J. Arsanjani. (2015). "Spatial Eigenvector Filtering for Spatiotemporal Crime Mapping and Spatial Crime Analysis." *Cartography and Geographic Information Science* 42(2), 134–48.

Hughes, J., and M. Haran. (2013). "Dimension Reduction and Alleviation of Confounding for Spatial Generalized Linear Mixed Models." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 75(1), 139–59.

Jones III, J. P., and E. Casetti (eds.) (1992). *Applications of the Expansion Method*. London: Routledge.

Krosnick, J. A., P. S. Visser, and A. L. Holbrook. (1998). "American Opinion on Global Warming: The Impact of the Fall 1997 Debate." *Resources* 133, 5–9.

Kwan, M.-P. (1998). "Space-Time and Integral Measures of Individual Accessibility: A Comparative Analysis Using a Point-Based Framework." *Geographical Analysis* 30(3), 191–216.

Kwan, M.-P. (2004). "GIS Methods in Time-Geographic Research: Geocomputation and Geovisualization of Human Activity Patterns." *Geografiska Annaler: Series B, Human Geography* 86(4), 267–80.

Kwan, M.-P., and T. Neutens. (2014). "Guest Editorial: Space-Time Research in GIScience." *International Journal of Geographical Information Science* 28(5), 851–54.

Langran, G., and N. Chrisman. (1988). "A Framework for Temporal Geographic Information." *Cartographica* 25, 1–14.

LeSage, J. P., and M. M. Fischer. (2008). "Spatial Growth Regressions: Model Specification, Estimation and Interpretation." *Spatial Economic Analysis* 3(3), 275–304.

Liao, F., S. Rasouli, and H. Timmermans. (2014). "Incorporating Activity-Travel Time Uncertainty and Stochastic Space–Time Prisms in Multistate Supernetworks for Activity-Travel Scheduling." *International Journal of Geographical Information Science* 28(5), 928–45.

Lusher, D. (2013). "IP Addresses: Exploring the Nature of the Geographic Data and the Patterns that can be Extracted, Thesis. San Diego State University.

March, H., D. Saurí, and J. Olcina. (2014). "Rising Temperatures and Dwindling Water Supplies? Perception of Climate Change Among Residents of the Spanish Mediterranean Tourist Coastal Areas." *Environmental Management* 53, 181–93.

McCright, A. M., and R. E. Dunlap. (2011). "The Politicization of Climate Change and Polarization in the American Public's Views of Global Warming, 2001-2010." *Sociological Quarterly* 52(2), 155–94.

Nagel, A. C., M.-H. Tsou, B. H. Spitzberg, L. An, J. M. Gawron, D. K. Gupta, J.-A. Yang, S. Han, K. M. Peddecord, S. Lindsay, and M. H. Sawyer. (2013). "The Complex Relationship of Realspace Events and Messages in Cyberspace: Case Study of Influenza and Pertussis Using Tweets." *Journal of Medical Internet Research* 15(10), e237.

Pace, R. K., J. P. Lesage, and S. Zhu. (2013). "Interpretation and Computation of Estimates from Regression Models Using Spatial Filtering." *Spatial Economic Analysis* 8(3), 352–69.

Peuquet, D. J. (1994). "It's About Time: A Conceptual Framework for the Representation of Temporal Dynamics in Geographic Information Systems." *Annals of the Association of American Geographers* 84(3), 441–61.

Peuquet, D. J., and N. Duan. (1995). "An Event-Based Spatiotemporal Data Model (ESTDM) for Temporal Analysis of Geographical Data." *International Journal of Geographical Information Systems* 9(1), 7–24.

Preacher, K. J., A. L. Wichman, R. C. MacCallum, and N. E. Briggs. (2008). *Latent Growth Curve Modeling*. Los Angeles: SAGE Publications.

Rey, S. (2001). "Spatial Empirics for Economic Growth and Convergence." *Geographical Analysis* 33, 195–214.

Rey, S. J., and M. V. Janikas. (2006). "STARS: Space–Time Analysis of Regional Systems." *Geographical Analysis* 38(1), 67–86.

Scannell, L., and R. Gifford. (2013). "Personally Relevant Climate Change: The Role of Place Attachment and Local Versus Global Message Framing in Engagement." *Environment and Behavior* 45(1), 60–85.

Shavitt, Y., and N. Zilberman. (2010). "Study of Geolocation Databases." *Cornell University Library*. Online at http://arxiv.org/abs/1005.5674.

Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis*. New York: Chapman and Hall.

Smith, T. E. (2009). "Estimation Bias in Spatial Models with Strongly Connected Weight Matrices." *Geographical Analysis* 41(3), 307–32.

Spitzberg, B. H. (2014). "Toward a Model of Meme Diffusion (M^3D)." *Communication Theory* 24, 311–39.

Tiefelsdorf, M., and D. A. Griffith. (2007). "Semiparametric Filtering of Spatial Autocorrelation: The Eigenvector Approach." *Environment and Planning A* 39(5), 1193–221.

Tobler, W. R. (1970). "A Computer Moving Simulating Urban Growth in the Detroit Region." *Economic Geography* 46, 234–40.

Tsou, M.-H., I.-H. Kim, S. M. Wandersee, D. Lusher, L. An, B. H. Spitzberg, D. K. Gupta, J. M. Gawron, J. Smith, J.-A. Yang, and S. Y. Han. (2013). "Mapping Ideas from Cyberspace to Realspace: Visualizing the Spatial Context of Keywords from Web Page Search Results." *International Journal of Digital Earth* 7(4), 316–35.

Tsou, M.-H., and M. Leitner. (2013). "Editorial: Visualization of Social Media: Seeing a Mirage or a Message?" *Cartography and Geographic Information Science* 40(2), 55–60.

Worboys, M. F. (1992). "A Model for Spatio-Temporal Information." In *Proceedings: The 5th International Symposium in Spatial Data Handling*, 602–611. San Jose, CA: American Congress of Surveying and Mapping.

Wu, B., R. Li, and B. Huang. (2014). "A Geographically and Temporally Weighted Autoregressive Model with Application to Housing Prices." *International Journal of Geographical Information Science* 28(5), 1186–204.

Yuan, M. (1999). "Use of a Three-Domain Representation to Enhance GIS Support for Complex Spatiotemporal Queries." *Transactions in GIS* 3(2), 137–59.

Yuan, M., A. Nara, and J. Bothwell. (2014). "Space-Time Representation and Analytics." *Annals of GIS* 20(1), 1–9.