Understanding information landscape through space-time analysis

Li An, lan@mail.sdsu.edu, Department of Geography, San Diego State University

Analysis of cyberspace data is a rapidly growing research frontier in many disciplines including geography, political science, and epidemiology. Recent advances in information science and GIScience technologies, along with the increasing availability of large space-time datasets ("Big data"), have engaged scientists of varying background in space-time analysis and modeling of cyberspace data. In this context, we have chosen two exemplar keywords of "climate change" and "global warming" to collect data through web search (based on the Yahoo search engine) on a weekly basis. Accordingly, we have developed a set of methods to verify and to analyze these data in the hope that cyberspace data can be used to reveal realspace processes. Our goal is thus two folds. First, we are interested in knowing more about the usage domain of web search data including the linkage between cyberspace and realspace data; second, we aim to establish a space-time analysis methodology that may be broadly applicable.

In regards to the usage domain of web search data, our analysis results reveal that 1) around 25-30% of websites have their IP registration address "correctly" reflecting their real world address, i.e., website offices/company headquarters is not in the same city as the server (i.e., these sites are free of the dislocation problem); 2) the rank (based on popularity) to which a website has been assigned (on a certain topic such as "climate change" and "global warming") can largely be used to represent the interest of real world people near the website on the topic if the website is geolocationally or otherwise verified; 3) GIS krigging and kernel density mapping methods can be used to create the corresponding information surface for different (e.g., visualization, data analysis) purposes; 4) classification of websites could substantially filter out "noises" in web search data, and the classified data can better represent realspace phenomena (here interest in a certain topic from a certain targeted real world group). The education websites and government websites (URL ending with edu or gov) out of the twelve web classes we have identified are less prone to the dislocation problem; 5) even without the dislocation problem, there may exist a gap between cyberspace data and real space phenomena due to biases arising from many factors (e.g., who are more likely to use websites to express what kind of opinions). Related to our second goal, we have done the following for space-time analysis.

1) Data collection and preparation: We assembled a large amount of real space data for later regression analysis. The sources include American Community Survey (2005-09), North America Climate Extremes Monitoring (2006), US Census (2010), US Agricultural Census (2007), National Agricultural Statistics Service (2010), and National Weather Service Storm Prediction Center (2011). In order to automate the processing of keyword search results, we developed R scripts (for data cleaning and formatting) and several models in ArcGIS using Python to facilitate batch processing to handle the large time series of weekly data.

2) Metrics development: Several metrics were employed to quantify spatial patterns at discrete times, including those used in landscape ecology (e.g., dominance, proximity), map comparison (e.g., Pontius et al. 2004, 2011), and spatial statistics (e.g., univariate and bivariate LISA, Geary's G; Anselin et al. 2006). Specifically, we have pursued exploratory spatial analysis in GeoDa, including use of spatial lag and error models, univariate and bivariate local indicators of spatial association.

In particular, we used the "hazard" (a term quite often used in sociology, demography, and epidemiology) concept to depict and quantify the risk of an area being dominated or substantially influenced by certain events or ideas. Although in some instances hazard is equal to probability numerically, hazard is quite different from probability conceptually. As a sort of intensity measure, the

hazard of some event can take any nonnegative (i.e., not necessarily less than 1) value at each specific time point, independent of whether we can observe them or not. The probability of the event, ranging from 0 to 1, has to be defined and calculated within an appropriate time interval, i.e., $[t, t + \Delta t]$ by definition. For convenience of understanding, we can think of the hazard as the number of events expected within a short interval of time if the hazard is constant within this interval. The concept hazard could sometimes be simplified as a function of survival time, which is the time within which a place is free from a certain event (Allison 1995, p.63-66; An and Brown 2008).

3) Multivariate regression: We applied multivariate regression (e.g., OLS, binomial or multinomial logistic regression) in SAS in order to link realspace data (biophysical, socioeconomic, and demographic characteristics of the realspace) and cyberspace data related to reflections, attitudes, or perceptions towards "climate change" and "global warming". Survival analysis has been heavily used in this regard, in which hazards (of a certain place being "dominated" by interest in climate change or global warming) were linked with a set of independent variables. Survival analysis is known to be especially appropriate to deal with independent variables taking changing values over time or data related to imprecise time stamps for events. We found that 1) climatic patterns (amount of precipitation, summer days, etc.) significantly predict the Internet volume for "climate change" and "global warming"; 2) socioeconomic and demographic features, such as age and ethnicity, all have significant influences on such reflections or perceptions; 3) people's attitudes towards these terms may be partially explained by the dominant political party membership of a certain region; and 4) "global warming" as a term returns more negative and neutral websites and expired websites. We used various model correctness or fit measures (e.g., R^2 or pseudo R^2, AIC) to evaluate the models. The R^2 varies from 0.30 to 0.60.

4) On the other hand, we are also (maybe more) interested in using the regression in a non-tradition manner. As mentioned earlier, there may be biases in cyberspace data when we want to infer realspace patterns. Under certain social science or information science theories, we may use real space data in regression and get some reasonable regression results. When we replace real space data with related cyber space data (unable to completely verify such data's reliability or completeness) and get similar reasonable regression results, we may increase our confidence in the corresponding cyber space data. Our use of this method (i.e., non-traditional use of regression) has generated some great consistent results in terms of verifying the reliability or completeness of cyber space data.

Acknowledgement: This position paper is based on the NSF CDI project "Mapping cyberspace to real-space: visualizing and understanding the spatiotemporal dynamics of global diffusion of ideas and the semantic web". Thanks go to Drs. M. Tsou (PI), Dipak K. Gupta , Jean Mark Gawron, and Brian Spitzberg (co-PIs) and students.

References

1. Allison, P. D. 1995. Survival analysis using SAS: A practical guide. Cary, NC: SAS Institute.
2. An, L., and D. G. Brown. 2008. Survival analysis in land-change science: integrating with GIScience to address temporal complexities. Annals of Association of American Geographers, 98(2): 323-344.
3. An, L., D. G. Brown, J. 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.
4. Anselin, L., Syabri, I., and Kho, Y. 2006. GeoDa: An introduction to spatial data analysis. Geographical Analysis, 38(1):5–22.
5. Pontius, R.G., Jr., E. Shusas, and M. McEachern. 2004. Detecting important categorical land changes while accounting for persistence. Agriculture, Ecosystems and Environment, 101: 251–268.
6. Pontius Jr, RG, S Peethambaram, and J-C Castella. 2011. Comparison of three maps at multiple resolutions: a case study of land change simulation in Cho Don District, Vietnam. Annals of the Association of American Geographers 101(1): 45-62.