A High-Throughput Computational Approach to Environmental Health Study Based on CyberGIS

Xun Shi[1], Shaowen Wang[2], and Anand Padmanabhan[2]

CyberGIS has emerged as a fundamentally new modality of geographic information systems (GIS). Enabled by advanced cyberinfrastructure, it represents a holistic approach to combining data-intensive and computational sciences for geographic problem solving (Wang 2010; Wang *et al.* 2013). By taking advantage of modern network and computing technologies, it tackles complex geographic problems on which conventional GIS lack capability. Specifically, cyberGIS software environment enables sciences by 1) providing an online environment for making CyberGIS capabilities accessible to a large of number of users for research and educational purposes, through the CyberGIS Gateway; 2) enabling loosely coupled scalable geospatial software capabilities within advanced cyberinfrastructure environments such as the NSF Extreme Science and Engineering Discovery Environment (XSEDE) and Open Science Grid, via the CyberGIS Toolkit; and 3) managing the complexity of interacting with heterogeneous and distributed resources and services of advanced cyberinfrastructure through standard and friendly interfaces, provided by GISolve middleware. In this paper, we focus on implementing a computational approach to geospatial analysis for environmental health study based on cyberGIS.

The approach relies on computation on three aspects: 1) disaggregating aggregate data; 2) calculating local statistics; and 3) evaluating statistical significance. Under this framework, we have developed methodologies for fine scale disease mapping and for disease-environment association evaluation.

Aggregate location data is common in environmental health studies. Disaggregating such data may improve the quality of analysis and quantify the uncertainty caused by data aggregation (Shi). We developed a *restricted and controlled Monte Carlo* (RCMC) process to disaggregate areal level location data. RCMC allocates polygon-level locations of the subjects, e.g., patients, to random point locations. The randomization is restricted by the areal unit to which a subject belongs, i.e., the subject cannot be assigned to a location outside its own unit. The randomization is also controlled by high-resolution population information, i.e., the chance for a place (represented by a raster cell) to receive a subject is proportional to the population in that place. The randomization is repeated many times, and the variability in the results from these many randomizations represents the uncertainty caused by data aggregation.

Once all location data are at the point level, for mapping disease, we use kernel density estimation (KDE) to calculate local disease intensity. Different from the conventional KDE implemented in most of the commercial off-the-shelf GIS tools, KDE in disease mapping needs to take into account of the *background* (e.g., the population at risk), i.e., this is KDE over an inhomogeneous background (Shi 2010).

We then use an *unrestricted but controlled Monte Carlo* (UCMC) process to evaluate statistical significance of a local intensity value. Specifically, subjects will be assigned to random locations in the study area; this randomization is no longer restricted by the areal units, but still controlled by the high-resolution population data. The intensity calculated through KDE based on the points generated in this way is considered an H_0 scenario. Many H_0 scenarios generated in this way can be used to evaluate the *p*-value of the intensity value generated from RCMC.

Still using the disaggregated location data from RCMC, for evaluating association between a disease and a certain environmental factor, we read environmental exposure value for each case or control. We then perform statistical test around each location in the study area to evaluate if the exposure values of cases are significantly different from those of controls around that location. The result is a *map of significance* showing spatial variation of the association between the disease and the environmental factor.

This approach is featured by Monte Carlo processes and local calculation, and thus critically relies on the computing capacity of the platform it is implemented on. The conventional GIS sets serious limitation to the adoption of this approach in public health studies and practice, although the approach is scientifically sound and advantageous, and is relative easy to understand and implement, compared with those statistically intensive methods. This kind of limitation is exactly what cyberGIS aims to resolve. We have started the migration of this approach from the conventional PC environment to a cyberGIS platform.

1. Department of Geography, Dartmouth College.
2. Department of Geography, University of Illinois at Urbana-Champaign.

We aim to allow the advantages of the approach to be fully realized by providing researchers and practitioners with seamless access to powerful cyberinfrastructure resources.

References:

Shi X, 2007, Evaluating the Uncertainty Caused by P.O.Box Addresses in Environmental Health Studies: A restricted Monte Carlo Approach. *International Journal of Geographical Information Science*, 21(3):325–340.

Shi X, 2009, A GeoComputation Process for Characterizing the Spatial Pattern of Lung Cancer Incidence in New Hampshire. *Annals of the Association of American Geographers*, 99(3):521–533.

Shi X, 2010, Selection of Bandwidth Type and Adjustment Side in Kernel Density Estimation over Inhomogeneous Backgrounds. *International Journal of Geographical Information Science*, 24(5):643–660.

Wang, S. 2010. "A CyberGIS Framework for the Synthesis of Cyberinfrastructure, GIS, and Spatial Analysis." *Annals of the Association of American Geographers*, 100(3): 535-557

Wang, S., Anselin, L., Bhaduri, B., Crosby, C., Goodchild, M. F., Liu, Y., and Nyerges, T. L. "CyberGIS Software: A Synthetic Review and Integration Roadmap*." International Journal of Geographical Information Science*, DOI:10.1080/13658816.2013.776049.

1. Department of Geography, Dartmouth College.
2. Department of Geography, University of Illinois at Urbana-Champaign.