Applied CyberGIS in the Age of Complex Spatial Health Data

Marta Jankowska: University of California San Diego

Advances in data acquisition in a number of fields throughout the health spectrum are resulting in large, complex, and diverse data sets. With proliferation of sensor and spatial data acquisition, the analysis and processing of complex spatial health data analytics will become a pressing problem. CyberGIS can offer solutions in this realm, however no systems have been developed that cater to the specific challenges associated with complex spatial health data such as privacy, real-time analytics, data standardization, data integration, workflow provenance, and a front end interface that is accessible to individuals in the public health realm. During the 2016 NSF-IBSS workshop, I would like to discuss and gain feedback on SPACES, an in-development CyberGIS with the focus of addressing some of these challenges.

Public health, health care, and medical advances are increasingly looking to the collection of voluminous data sets. Movements such as Quantified Self, where individuals engage in self-tracking of biological, physical, behavioural, and environmental information, or the Precision Medicine Initiative Cohort announced by President Obama in 2015, which will include extensive data collection and tracking of over one million U.S. participants, are driving the need to develop cyberinfrastructures and methodologies for big and complex spatial health data integration, processing, and analysis. Complex spatial health data may include temporally-linked objective measures of behavior and health, molecular data such as genomics, proteomics, or metabolomics, life course data including movement trajectories, self-reported survey and demographic data, environmental data assessing exposures, and finely resolved spatial data.

CyberGIS may assist in addressing the challenges of complex spatial health data. It represents a new-generation GIS based on the synthesis of advanced cyberinfrastructure, geographic information science, and spatial analysis and modeling. However, advances in the application of CyberGIS to health specific problems are lacking. Goldberg *et al.* presented a comprehensive vision of a Spatial-Health CyberGIS Marketplace, which tackled many of the opportunities as well as challenges of a health focused CyberGIS including confidentiality and privacy protections, real-time analytic methods, data standardization, and a comprehensive end-to-end ecosystem architecture. I would add to this list the need for shareable workflows to promote inter-field collaboration, diverse data type integration, and replicability of analytic processes.

Recently, the NSF funded a platform called DELPHI (Data E-platform for personalized population health), one of the first cyberinfrastructures in development specifically catered toward meeting the goals of integrating complex and diverse health data with an accessible API for individual use, clinician intervention and patient management, and population-level research. However, DELPHI has not been developed to operate in a HIPAA-compliant environment (HIPAA Privacy Rule establishes national standards for protection individual medical records and personal health information), it does not cater to the complexities of performing analyses on spatial data, and finally it does not offer workflows to promote collaboration and replication of processing and analytic tasks. For this reason, we have integrated DELPHI into a larger CyberGIS called SPACES (Secure Physical ACtivity and Environment Software).

SPACES is an in-development CyberGIS specifically geared at solving and promoting data integration, analysis, and collaboration with complex spatial health data. SPACES aims to address four core problems in existing CyberGIS applications for health: 1) conceptual and applied challenges of complex spatial and biomedical data integration, data integrity, and common data standards, 2) lack of easy-to-use data integration computation infrastructure, 3) lack of documented and shared workflows that will promote scalability, provenance (understanding of origin of results and repeatability of scientific process), and

knowledge base development, and 4) lack of systems designed to function in HIPAA-compliant environments.

Here, I focus on one specific application of complex spatial health data. The current obesity and inactivity epidemics have instigated a surge of research into the spatial factors that influence physical inactivity and obesity. Technological and methodological developments have led to the ability to examine dynamic, high-resolution measures of temporally-matched location and physical activity behavior data through GPS, accelerometry, and GIS, a nascent field called 'spatial energetics'. This field sees two promising paths by improving population health through environmental modifications and improving individual health through targeted mobile-based health interventions. Spatial energetics offers an ideal case study for the development of CyberGIS for complex health data by offering different types of data collection methods (real-time, data logging, individual self-report, survey), diverse levels of data collection (biological, minute level, individual demographic, community environment), large data sets (data sampled at the sub-minute level on large populations quickly grows to terabytes of data), and often focuses on health outcomes beyond obesity such as cancer or diabetes that requires HIPAA privacy protections.

SPACES is designed to support and train researchers on the spatial and temporal analysis of large volumes of physical activity, geographic, and contextual data. It is housed in a HIPAA and FISMA-compliant, private computing cloud housed at the San Diego Supercomputer Center called Sherlock; however the structure can be replicated in other HIPAA compliant clouds. The key organizing principle of SPACES is workflows, and a unified data communication layer through which the development of common data elements can be achieved. The workflows are provided by Kepler Scientific Workflow System, which provides a graphical user interface for designing workflows composed of a linked set of extensible and configurable components. Kepler can also call on distributed programming that supports interfacing to different components of the cyberinfrastructure and computing platforms. In SPACES, Kepler workflows are a programming abstraction that describes which computational elements (serial or parallel) and data elements need to fit together, and what order they must be processed in. In Figure 1 below, an example of such a workflow is given with processing and analysis blocks working through the integration of minute level accelerometer, GPS, and heart rate data.

Sitting below the workflow(s) are the systems and software deployments required for reliable execution, as well as the data communication layer for efficient data organization with is built upon the existing DELPHI infrastructure and adds in a PostgreSQL geodatabase for spatial data. A key element of SPACES is the creation of common data elements through existing physical activity and spatial energetics processing algorithms such as the Personal Activity Location Measurement System, machine learning algorithms, and spatial element extraction and analysis. Common data elements and outputs can directly enhance and influence the training efforts for the health research community and promote common data organization standards. The use of the Sherlock environment as a platform (or another HIPAA compliant platform) where all data integration and analysis takes place ensures security of sensitive data while allowing access to users to analyze derived datasets without compromising the raw data. Finally, a friendly user front end with training modules and ability to share workflows will be developed to make SPACES accessible to various researchers and end users.