Geospatial CyberInfrastructure in Land Use/Cover Change Research

Jin Xing and Renee Sieber

McGill University, Montreal, Quebec, Canada
Email: jin.xing@mail.mcgill.ca; renee.sieber@mcgill.ca

Keywords: CyberGIS, Domain Knowledge, Geospatial CyberInfrastructure, Big Data, Land Use/Cover Change.

1. Domain Knowledge Integration in LUCC-GCI

Detecting land use changes (e.g., the transition of forest to urban) should be made easier with big data but it is not. Growing data volumes allows us to observe Land Use/Cover Change (LUCC) at hyper-resolutions. However, research suggests that phenomena exist at bounded space-time granularities and extents (Vance and Doel, 2010). Therefore, increasing granularity likely will result in misclassified LUCC. Big data affords greater variety of, for example, available spectra and file types. Each LUCC data format may require a different change detection algorithm and data processing tool. Big velocity exacerbates volume and variety and can render LUCC more difficult to ground truth. We may not easily know when a forest has changed to urban or if we are observing impervious surfaces amid the trees. Consequently we are interested in the use of Geospatial CyberInfrastructures (GCIs) (Wang et al., 2013) to allow us to integrate domain knowledge in big data-enabled LUCC.

In prior research (Xing and Sieber, 2016), we created a GCI for LUCC. We used temporal topology information and image segmentation to decompose/recompose big data and detect changes. A drawback of this approach is these temporal topology rules are domain-independent. To account for domain knowledge, we manually examined LUCC big data results, via sampling, to guarantee the designation of LUCC types, the effectiveness of change detection algorithms (e.g., Radke et al., 2005), and accurate boundaries between "change" and "no-change" for different types of LUCC. This was the extent of our domain knowledge verification. Manual domain knowledge verification can no longer match the sheer volume, variety, and velocity of big data, and ensure new knowledge is discovered in the process of LUCC. Sampling cannot cover all the information for domain knowledge, according to the rule of "population, not sampling" (Miller and Goodchild, 2015). Thus the domain knowledge verification must rely on automation and a tighter coupling with GCI.

We argue for a hybrid parallel-serial workflow to integrate domain knowledge. The other layers of the GCI (see below) are parallelized. Domain knowledge integration in the GCI becomes a recursive process that involves data, the analysis methods, and the expertise of research scientists. Domain scientists provide the initial knowledge, by supplying components of the layer with LUCC labels, the change detection algorithms, other domain-based initializations to GCI. The LUCC workflow is executed and intermediate results update the existing domain knowledge base. Updates for LUCC include adding or deleting the LUCC labels, selecting different change identification algorithms, tuning parameters, and improving the workflow and resource allocation. The process continues until we cannot find new types of LUCC from the data, or the accuracy of the change identification exceeds pre-defined thresholds, or the boundaries of different LUCC areas converge. GCI should allow us to complete the domain knowledge verification process.

2. Design and Deployment

We propose a LUCC-GCI composed of five layers: the domain layer, the LUCC layer, the resource allocation layer, the dataflow management layer, and the computation layer. As suggested above, the domain layer provides instructions for the other four layers. At the beginning of each recursion, the domain layer checks the domain knowledge (e.g., the accuracy of feature extraction, LUCC label, and “change” area boundaries) from the last recursion, and implements corresponding updates. Initial domain knowledge plays a pivotal role in the GCI because inadequate initialization of the domain knowledge will create lags in functioning of the GCI. The domain knowledge is updated through the recursive LUCC workflow, until reaching the max number of recursion or achieving satisfactory results (according to domain-based fitness rules). The LUCC layer hosts the LUCC detection analysis code, including pre-processing, image change detection algorithms, and accuracy assessment methods. The resource allocation layer employs the spatial domain computation representation (Wang and Armstrong, 2009) to allocate the computing resource for different tasks. The dataflow management layer implements decomposition and recomposition methods for big data analysis (Xing, Sieber, and Kalacska, 2014). Finally, the computation layer integrates cloud computing and a Hadoop framework to provide scalable resource provisioning, and data storage management (Almeer, 2012).

We evaluate our change detection GCI with the urban-rural change detection in the Greater Montreal Area (2006-2012) covering 36,000km^2. The GCI is built on top of the private-public cloud composed of our local computing nodes and Microsoft Azure resources. Result suggest the potential for domain knowledge discovery in big data-enabled LUCC.

Acknowledgement

This research is funded by Microsoft Azure Research Award.

References

Almeer, M. H. (2012). Cloud hadoop map reduce for remote sensing image analysis. *Journal of Emerging Trends in Computing and Information Sciences*, 3(4), 637-644.

Kitchin, R. (2013). Big data and human geography Opportunities, challenges and risks. *Dialogues in Human Geography*, 3(3), 262-267.

Miller, H. J., & Goodchild, M. F. (2015). Data-driven geography. *GeoJournal*, 80(4), 449-461.

Radke, R. J., Andra, S., Al-Kofahi, O., & Roysam, B. (2005). Image change detection algorithms: a systematic survey. *IEEE Transactions on Image Processing*, 14(3), 294-307.

Vance, T. C., & Doel, R. E. (2010). Graphical methods and Cold War scientific practice: the Stommel diagram's intriguing journey from the physical to the biological environmental sciences. *Historical Studies in the Natural Sciences* , 40(1), 1-47.

Wang, S., Anselin, L., Bhaduri, B., Crosby, C., Goodchild, M. F., Liu, Y., & Nyerges, T. L. (2013). CyberGIS software: a synthetic review and integration roadmap. *International Journal of Geographical Information Science*, 27(11), 2122-2145.

Wang, S., & Armstrong, M. P. (2009). A theoretical approach to the use of cyberinfrastructure in geographical analysis. *International Journal of Geographical Information Science*, 23(2), 169-193

Xing, J., & Sieber, R. E. (2016). A land use/land cover change geospatial cyberinfrastructure to integrate big data and temporal topology. *International Journal of Geographical Information Science*, 30(3), 573-593.

Xing, J., Sieber, R., & Kalacska, M. (2014). The challenges of image segmentation in big remotely sensed imagery data. *Annals of GIS*, 20(4), 233-244.