International Cartographic Association
Association Cartographique Internationale

Towards the large-scale extraction of historical land cover information from historical maps

Johannes H. Uhl [a,b*], Stefan Leyk [b,c], Weiwei Duan [d], Zekun Li [d], Basel Shbita [e], Yao-Yi Chiang [d] and Craig A. Knoblock [e]

[a] *Earth Lab, Cooperative Institute for Research in Environmental Sciences (CIRES), University of Colorado Boulder, USA. Johannes.Uhl@colorado.edu*

[b] *Institute of Behavioral Science, University of Colorado Boulder, USA*

[c] *Department of Geography, University of Colorado Boulder, USA*

[d] *Spatial Sciences Institute, University of Southern California, USA*

[e] *Information Sciences Institute, University of Southern California, USA*

* Corresponding author

Keywords: Map processing, Information extraction, Historical maps, Geospatial data integration, Digital humanities.

Abstract:

Historical maps contain valuable information on settlements, land cover, and transportation infrastructure in the past. However, the information contained in such documents needs to be converted into analysis-ready data formats in order to be used for retrospective landscape assessments or to inform long-term land cover / land use models. The recent availability of historical topographic map archives as collections of georeferenced, digital raster data has catalysed the development of methods for information extraction from historical maps using advanced image processing or machine-learning based computer vision methods (Uhl & Duan 2021). Most of these studies, however, focus on relatively small datasets and thus, it remains unknown whether such methods are applicable and scalable for the extraction of historical geographic features across large (e.g., country-level) study areas and massive data volumes.

We designed a pipeline to obtain large amounts (N>50,000) of historical topographic maps from the United States Geological Survey (USGS) in order to test the extraction of geographic features such as historical forest extents and historical urban areas over large areas. We focused on large-scale maps (i.e., map scales 1:24,000 and 1:62,500). While some of these maps date back to the 1890s, forest cover is typically contained in maps created in the 1950s or later. Based on the available metadata, map sheets with specific characteristics (e.g., containing woodland tint or not), or from specific time periods can be selected. Hence, we were able to select the earliest map sheet per quadrangle that contains woodland tint, for forest extraction, or the earliest available map per quadrangle contained within a given area of interest, for urban area extraction.

In order to keep computational costs to a minimum, we use low-level image descriptors (e.g., color moments, Huang et al. 2010) to generalize the map content (i.e., aggregate the spectral data) within grid cells of 100x100 meters, which can then be used as input for the extraction of the features of interest. For example, forest areas are typically represented in green and can be selected through appropriate image processing methods, including RGB-thresholding methods, unsupervised color-space clustering or advanced semantic segmentation methods. This grid-based strategy also facilitates the integration of other gridded data (e.g., remote sensing data or derived layers) into the processing chain. Such data integration facilitates (a) narrowing down search areas for specific geographic features, and (b) quantitatively cross-comparing raw or extracted map content to related gridded datasets for quality checks, accuracy assessments, as well as change detection, consistently and seamlessly across large amounts of individual map sheets. Moreover, the framework is independent from the actual information extraction method that can be incorporated in flexible ways.

We are currently testing the performance of different extraction methods with respect to the extraction quality and efficiency. While first results are promising, we find that the variety of map designs, defining both the appearance and content of maps, and high levels of heterogeneity in the temporal reference of the map sheets pose substantial challenges for the seamless extraction of map content across large amounts of map sheets despite the availability of detailed metadata. We find that our approach to extract information from the USGS topographic map archive at a country-level scale overcomes some of these key challenges and produces acceptable results.

The contribution of this work is four-fold. 1) We provide a framework for the large-scale extraction of historical map content across large amounts of individual map sheets; 2) We raise awareness about the issues and challenges related to the heterogeneity in map design, map content, and temporal reference when massive volumes of map sheets are processed; 3) We use the extracted historical urban extents in combination with more recent, remote-sensing derived urban extents,

https://doi.org/10.5194/ica-abs-3-297-2021

allowing for the assessment of long-term urbanization trajectories at the city or metropolitan area level; 4) We provide extraction results to the public, such as forest extents across the conterminous United States dating back to the time period between 1940 and 1970. Such data are extremely valuable for long-term environmental studies which are typically based on remote sensing data and thus, constrained to more recent time periods.

References:

Uhl, J. H., & Duan, W. (2021). Automating Information Extraction from Large Historical Topographic Map Archives: New Opportunities and Challenges. In: Werner, M. and Chiang, Y.-Y. (Eds.): *Handbook of Big Geospatial Data*. Springer, Cham.

Huang, Z. C., Chan, P. P., Ng, W. W., & Yeung, D. S. (2010, July). Content-based image retrieval using color moment and Gabor texture feature. In *2010 International conference on machine learning and cybernetics* (Vol. 2, pp. 719-724). IEEE.

https://doi.org/10.5194/ica-abs-3-297-2021