 doi:10.20944/preprints201803.0021.v1

Article

Map Archive Mining: Visual-analytical Approaches to Explore Large Historical Map Collections

Johannes H. Uhl [1,*], Stefan Leyk [1], Yao-Yi Chiang [2], Weiwei Duan [2] and Craig A. Knoblock [2]

[1] Department of Geography, University of Colorado Boulder, Boulder, Colorado, USA; {johannes.uhl;stefan.leyk}@colorado.edu

[2] Spatial Sciences Institute, University of Southern California, Los Angeles, California, USA; {yaoyic;weiweidu;knoblock}@usc.edu

* Correspondence: johannes.uhl@colorado.edu; Tel.: +01-303-492-2631

Abstract: Historical maps constitute unique sources of retrospective geographic information. Recently, several map archives containing map series covering large spatial and temporal extents have been systematically scanned and made available to the public. The geographic information contained in such data archives allows extending geospatial analysis retrospectively beyond the era of digital cartography. However, given the large data volumes of such archives and the low graphical quality of older map sheets, the processes to extract geographic information need to be automated to the highest degree possible. In order to understand the salient characteristics, data quality variation, and potential challenges in large-scale information extraction tasks, preparatory analytical steps are required to efficiently assess spatio-temporal coverage, approximate map content, and spatial accuracy of such georeferenced map archives across different cartographic scales. Such preparatory steps are often neglected or ignored in the map processing literature but represent highly critical phases that lay the foundation for any subsequent computational analysis and recognition. In this contribution we demonstrate how such preparatory analyses can be conducted using classical analytical and cartographic techniques as well as visual-analytical data mining tools originating from machine learning and data science, exemplified for the United States Geological Survey topographic map and Sanborn fire insurance map archives.

Keywords: map processing; retrospective landscape analysis; visual data mining, image retrieval, low-level image descriptors, color moments, t-distributed stochastic neighborhood embedding, USGS topographic maps, Sanborn fire insurance maps

---

1. Introduction

Historical maps contain valuable information about the Earth's surface in the past. This information can provide a detailed understanding of the evolution of the landscape as well as the interactions between and the drivers of geographic phenomena, such as human-made structures (e.g., transportation networks, settlements), vegetated land cover (e.g., forests, grasslands) or hydrographic features (e.g., stream networks, water bodies). However, this spatial information is typically locked in scanned map images and needs to be extracted to get access to the geographic features of interest in machine readable data formats that can be imported into geospatial analysis environments.

Map processing, or information extraction from digital map documents, is a branch of document analysis that focuses on the development of methods for the extraction and recognition of information in scanned cartographic documents. Map processing is an interdisciplinary field that combines elements of computer vision, pattern recognition, geomatics, cartography, and machine learning. The main goal of map processing is to "unlock" relevant information from scanned map documents to provide this information in digital, machine-readable geospatial data formats as a means to preserve the information digitally and facilitate the use of these data for analytical purposes [1].

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 March 2018 doi:10.20944/preprints201803.0021.v1

Remotely sensed earth observation data from space and airborne sensors has been systematically acquired since the early 1970s and provides abundant information for the monitoring and assessment of geographic processes and how they interact over time. However, for the time periods prior to operational remote sensing technology, there is little (digital) information that can be used to document these processes. Thus, map processing often focuses on the development of information extraction methods from map documents or engineering drawings created prior to the era of remote sensing and digital cartography, thus expanding the temporal extent for carrying out geographic analyses and landscape assessments to more than 100 years in many countries.

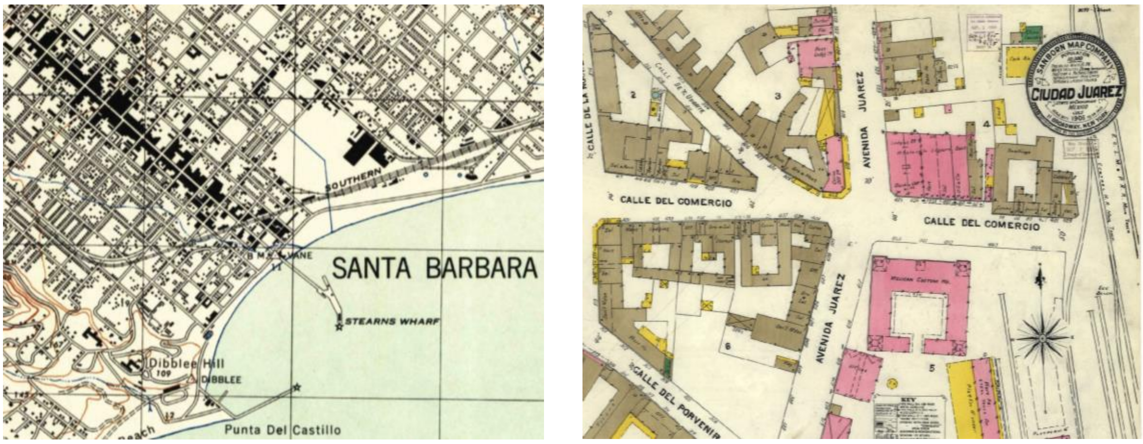
Information extraction from map documents includes the steps of *recognition* (i.e., identifying objects in a scanned map such as groups of contiguous pixels with homogeneous semantic meaning), and *extraction* i.e., transferring these objects into a machine-readable format (e.g., through vectorization). Extraction processes typically involve image segmentation techniques based on histogram analysis, color-space clustering, region growing or edge detection. Recognition in map processing is typically conducted using template matching techniques involving shape descriptors, cross-correlation measures or based on feature descriptors. Exemplary applications of map processing techniques include the extraction of buildings [2-4], road networks [5], contour lines [6], composite forest symbols [7] or the recognition of text from map documents [8,9]. Most approaches rely on handcrafted or manually collected templates of the cartographic symbol of interest and involve a significant level of user interaction, which impedes the application of such methods for large-scale information extraction tasks where high degrees of automation are necessary to process documents with high levels of variation in data quality.

The availability of abundant contemporary geospatial data for many regions of the world offers new opportunities to employ contemporary geospatial data as ancillary information to facilitate the extraction and analysis of geographic content from historical map documents. Such approaches include the use of contemporary spatial data for georeferencing historical maps [10], assessing the presence of objects in historical maps across time at locations dictated by contemporary geospatial vector data [11], or the automated collection of template graphics for cartographic symbols of interest based on locations derived from modern geospatial data sources [12].

Most existing approaches for content extraction from historical maps still require a certain degree of user interaction to ensure acceptable extraction performance for individual map sheets, e.g. [13]. To overcome this persistent limitation, [14] and [15] propose the use of active learning and similar interactive concepts for more efficient recognition of cartographic symbols in historical maps, whereas [16] examine the usefulness of crowd-sourcing for the same purpose.

Moreover, the recent developments in deep machine learning in computer vision and image recognition have catalyzed the use of such techniques for geospatial information extraction from earth observation data [17-26]. This methodological development naturally projects into the idea of applying state-of-the-art machine learning techniques for information extraction from scanned cartographic documents, despite their fundamentally different nature compared to remotely sensed data. Key in both cases is the need for abundant and representative training data which requires automated sampling techniques. First attempts in this direction have used ancillary geospatial data for the collection of large amounts of training data in historical maps [27-30].

Besides this, several efforts have recently been conducted in different countries to systematically scan, georeference, and publish whole series of topographic and other map documents. These developments include efforts at the United States Geological Survey (USGS), that scanned and georeferenced approx. 200,000 topographic maps published between 1884 and 2006 at different cartographic scales between 1:24,000 and 1:250,000 [31] and the Sanborn fire insurance map collection maintained by the U.S. Library of Congress, that contains more than 500,000 sheets of large-scale maps of approximately 12,000 cities and towns in the U.S., Canada, Mexico, and Cuba, out of which approximately 6,000 map sheets have been published as scanned map documents [32-34]. Figure 1 shows an example of a USGS topographic map sheet and a Sanborn map, respectively. Furthermore, the United Kingdom Ordnance Survey scanned and georeferenced more than 200,000 topographic map sheets and town plans for the United Kingdom dating back to the 1840s and provides many of them as seamless georeferenced raster layers [35,36].



(a) (b)

Figure 1. Examples of historical map documents: **(a)** USGS topographic map 1:31,680 from Santa Barbara (California, 1944) and **(b)** Sanborn fire insurance map from city center of Ciudad Juárez (Mexico, 1905).

These developments, alongside with advances in the processing, storage and distribution of large data volumes, offer great potential for automated, large-scale information extraction from historical cartographic document collections in order to preserve the contained geographic information and make it accessible for geospatial analysis. Because of the large amount of data contained in these map archives, high degrees of extraction automation are necessary. This constitutes a challenging task given the high variability in the content and quality of maps within an archive. Possible reasons for such variability are different conditions of the archived analogue map documents, differences in the scan quality, as well as changes in cartographic design best practices that may have resulted in different symbologies over multiple map editions (Figure 2).

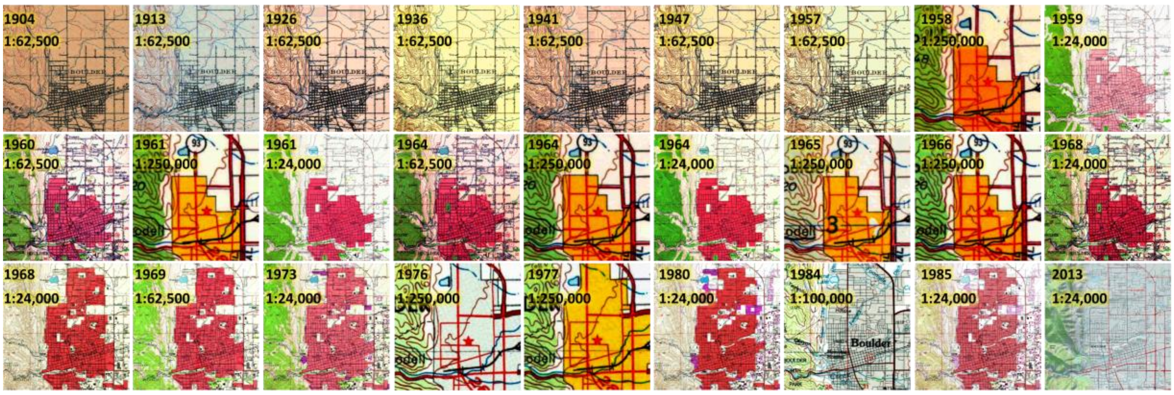


Figure 2. Example of the multi-temporal, multi-scale USGS topographic map archive, showing available map sheets covering Boulder, Colorado (USA) from 1904 to 2013 at various map scales.

The challenges described above summarize some of the central tasks in map archive processing which include dealing with the sheer data volume, the differences in cartographic scales and designs, changes in content and cartographic representations and their degree of similarity in individual maps, the spatial and temporal coverage of the map sheets, and the spatial accuracy of the georeferenced map which dictates the degree of spatial agreement to contemporary geospatial ancillary data. While the previously described approaches represent promising directions towards higher levels of automation, they imply that the graphical characteristics of the map content to be extracted are known and that map scale and cartographic design remain approximately the same

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 March 2018 doi:10.20944/preprints201803.0021.v1

across the processed map documents. Typically, many of these aspects are a priori unknown, since such large amounts of data cannot be analyzed manually. However, these are relevant pieces of information to better understand the data sources in order to design effective information extraction methods.

The remote sensing community faces similar challenges. The steadily increasing amount of remotely sensed earth observation data requires effective mining techniques to explore the content of large remote sensing data archives. Therefore, visual data mining techniques have successfully been used to comprehensively visualize the content of such archives. Such image information mining (IIM) systems allow to discover and retrieve using available metadata, and based on the similarity of the content of the individual datasets, or of patches of these [37-39] and guide exploratory analysis of large amounts of data which facilitates the subsequent development of information extraction methods. [40] implemented such a system for TerraSAR-X data, and [41] tested such approaches for patches of Landsat ETM+ data and the UC Merced benchmark dataset. These systems are based on spectral and textural descriptors precomputed at dataset or patch level that are then combined to multidimensional descriptors characterizing spectral-textural content of the datasets or patches. Other approaches include image segmentation methods to derive shape descriptors [42], include spatial relationships between images into the IIM [43], or make use of structural descriptors to characterize the change of geometric patterns over time across datasets within remote sensing data archives [44]. Comparison of these descriptors facilitates the retrieval of similar content across large archives. These approaches include methods for dimensionality reduction to visualize a whole archive in a two or three-dimensional feature space based on content similarity.

Whereas in remote sensing data archives the spatio-temporal coverage of the data and their quality is relatively well-known based on the sensor characteristics (e.g., time of operationality, satellite orbit, revisiting frequency, knowledge about physical parameters affecting data quality), this may not always be the case for historical map archives, where metadata on spatial-temporal data coverage might not be available or available in unstructured data formats only, preventing direct and systematic analysis.

Thus, there is an urgent demand to develop a systematic approach to explore such digital map archives, efficiently, prior to the extraction process, lending from similar efforts applied to remote sensing data, but with a stronger focus on information obtained by metadata analysis. In this contribution, we examine various techniques that could be used to build an image information mining system for digital cartographic document archives in combination with metadata analysis. These techniques allow to shed light on the following questions, which a potential user of such map archives may ask prior to the design and implementation of information extraction methods:

- **What is the spatial and temporal coverage of the map archive content and does it vary across different cartographic scales**? This is important because the coverage of the map data dictates the spatial and temporal extent of the information that can be extracted from the map archive, and thus, relates to the benefit of information extraction efforts and to the value of the extracted data. Furthermore, such an analysis is useful to compare different map archives.
- **How accurate is the georeference of the maps contained in the archive? Does accuracy vary in the spatio-temporal domain?** This constitutes a pressing question if ancillary geospatial data is used for the information extraction, which requires certain degrees of spatial alignment between map and ancillary data. For example, if it is possible to a-priori identify map sheets likely to suffer from a high degree of positional inaccuracy, the user can exclude those map sheets from template or training data collection, and thus, reduce the amount of noise in the collected training data.
- **How much variability is there in the map content, regarding color, hue, contrast, and in the cartographic styles used to represent the symbol of interest?** This is a central question affecting the choice and design of a suitable recognition model. More powerful models or even separate models for certain types of maps may be required if the representation of map content of interest varies heavily across the map archive. Furthermore, knowledge about variations in map content

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 March 2018 doi:10.20944/preprints201803.0021.v1

and about similarity between individual maps is useful regarding the training data sampling design, to ensure the collection of representative and balanced training samples.

We present a set of methods that illuminate these questions, based on metadata analysis and descriptor-based visual data mining. Systematic mining approaches of relevant information about the map archive help to inform and educate the user community on critical aspects of data availability, quality and spatio-temporal coverage. Furthermore, these exploratory steps provide insights that are relevant for the implementation of large-scale information extraction methods from historical map archives and help to anticipate potential challenges involved. These methods have proven to provide valuable information highly relevant to design information extraction methods presented in [27,28], e.g., regarding the choice of training areas and classification methods. These methods can be generalized to other existing map archives in similar ways as well. Additionally, we aim to raise awareness about the importance of a-priori knowledge on large spatial data archives before using the data for information extraction purposes. Such a preprocessing stage is often neglected in published research that traditionally focuses on the extraction methods, specifically. However, this is important, non-trivial work highly relevant in the age of data intensive research on information extraction. We exemplify these methods using the USGS topographic map archive and the Sanborn fire insurance map collection.

2. Data, Methods and Results

In this contribution, we propose a set of methods that can be used to explore the spatial-temporal coverage of a historical map archive, its content, existing variations in cartographic design and to partially assess the spatial accuracy of the maps. The approaches range from pure metadata analysis to descriptor-based visual data mining techniques. ***Metadata analysis*** is conducted for the USGS topographic map archive exemplified for the states of California and Colorado (USA) based on structured metadata, as well as for the Sanborn fire insurance map archive in the United States based on unstructured metadata. ***Content-based analysis*** is demonstrated for the USGS topographic map archive covering the state of Colorado at different map scales, involving the use of image descriptors, dimensionality reduction, data visualization methods, and similarity assessment based on geospatial ancillary data. The USGS map archive includes 14,831 map documents in California, and 6,964 map sheets in Colorado,

Both metadata analysis and content-based analysis constitute preparatory steps yielding valuable information that facilitates the design and implementation of information extraction methods based on large map archives. Figure 3 shows how the proposed methods can be incorporated in information extraction workflows.

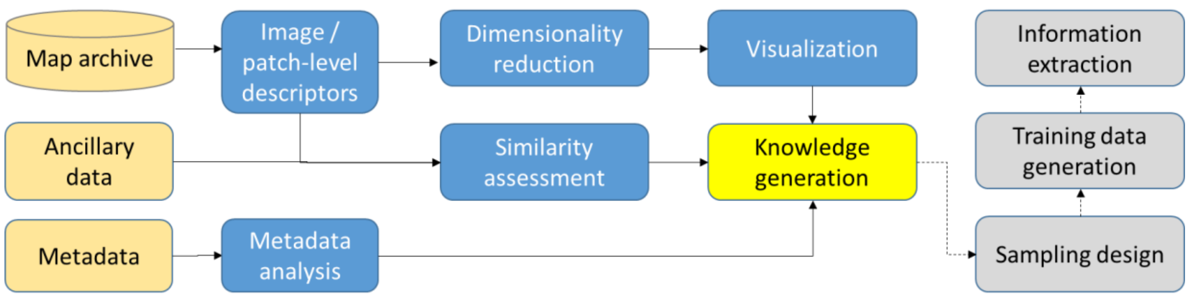


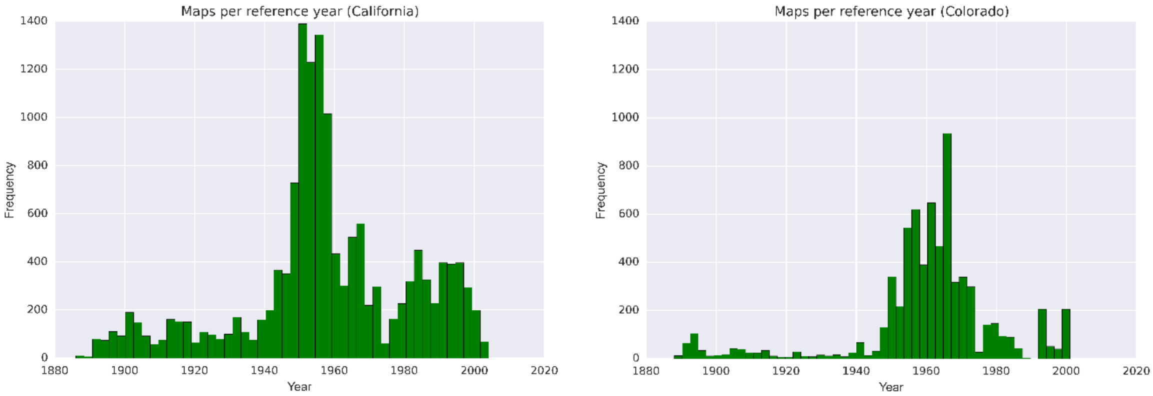
Figure 3. The methodology for metadata analysis and content-based knowledge generation on map archives to facilitate information extraction.

2.1. Metadata analysis

doi:10.20944/preprints201803.0021.v1

2.1.1. Metadata-based spatial-temporal coverage analysis

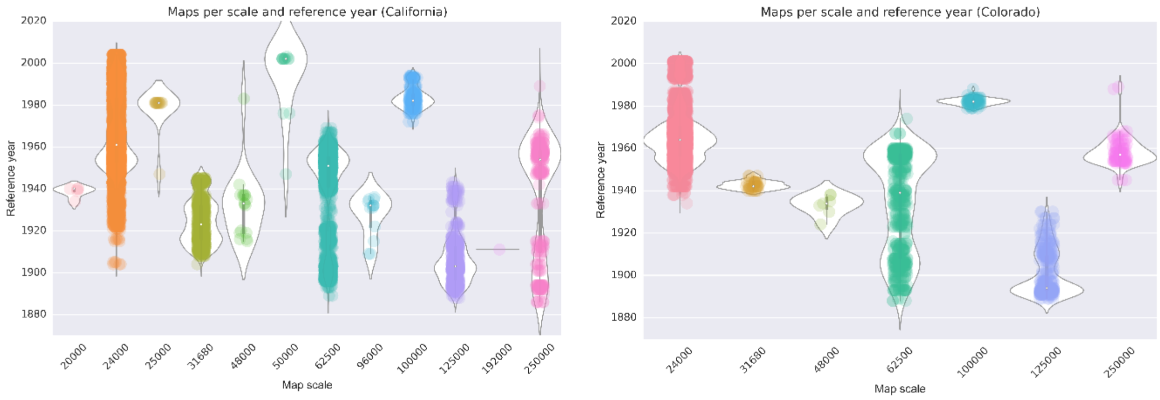
First, the temporal coverage of the map archives is analyzed. For the USGS map archive, which is accompanied by structured metadata (i.e., text files including unique identifiers for each map document), histograms based on the map reference year are created (Figure 4). It can be seen that the peak of map production was in California in the 1950s, and slightly later, in the 1960s in Colorado.



(a) (b)

Figure 4. Histograms of USGS topographic maps (all available map scales) by reference year, **(a)** in California, and **(b)** in Colorado (USA).

In addition to that, map production over time can be assessed in strata of map scales shown herein for the states of California and Colorado (Figure 5). These plots show the temporal distribution of published map editions (represented by the dots) and give an estimate of the underlying probability density (represented by the white areas) that indicates the map production intensity over time, separate and relative for each map scale. Such a representation helps to understand which time span can be covered with maps of various scales and thus can be used to determine which products to focus on for a particular purpose. This is important because maps of different scale contain different levels of detail resulting from cartographic generalization.



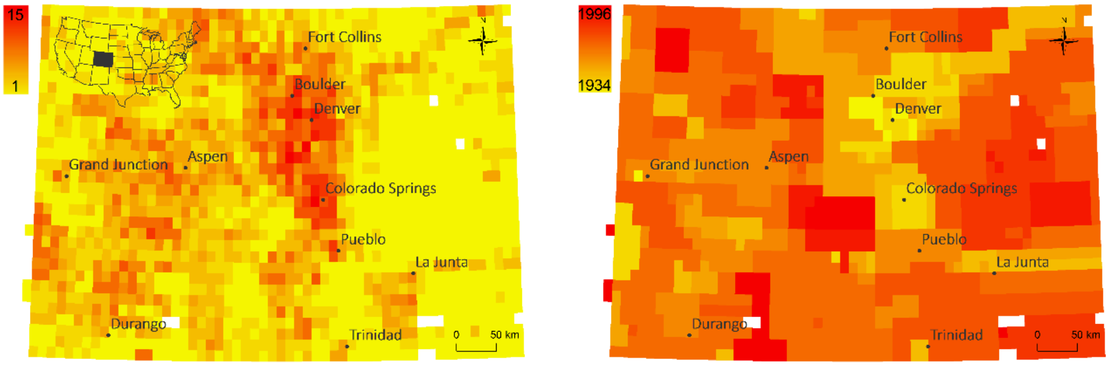
(a) (b)

Figure 5. Produced USGS topographic maps per reference year and map scale **(a)** in California, and **(b)** in Colorado (USA).

In order to assess the spatial variability of map availability in a map archive over time, spatial map footprints (i.e., delineating map quadrangles) are generated based on USGS-delivered metadata. For each map footprint, statistics about available map sheets at those locations are computed. This allows the spatial visualization of the number of map editions and the earliest reference year available for each location, as shown in Figure 6 for the state of Colorado (scale 1:24,000), and for the map scales 1:24,000 and 1:62,500 for the state of California in Figure 7, respectively. As can be seen, such

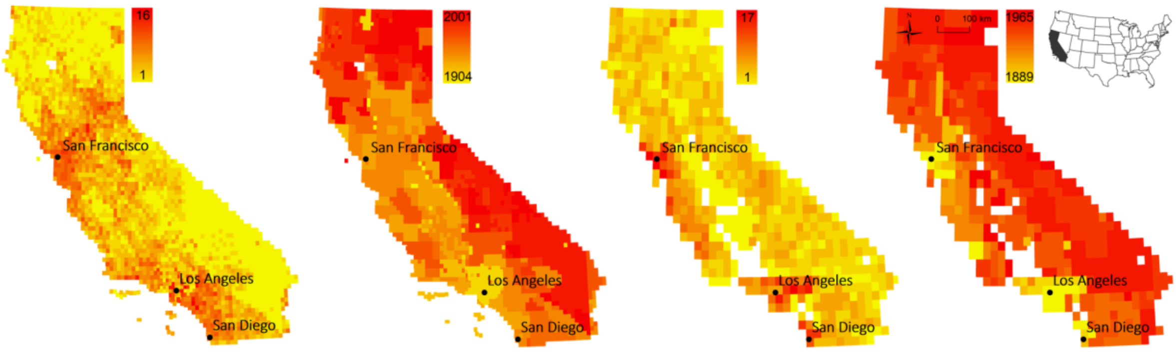
Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 March 2018 doi:10.20944/preprints201803.0021.v1

representations are useful to identify regions that have been mapped more intensively versus those for which temporal coverage is rather sparse. Furthermore, a user is immediately informed about the earliest map sheets for a location of interest to understand the maximum time period covered by these cartographic documents. Similar representations could be created for the average number of years between editions or the time span covered by map editions of a given map scale.



(a) (b)

Figure 6. (a) Map edition counts and **(b)** earliest map production year per 1:24,000 map quadrangle in the state of Colorado (USA) based on metadata analysis.

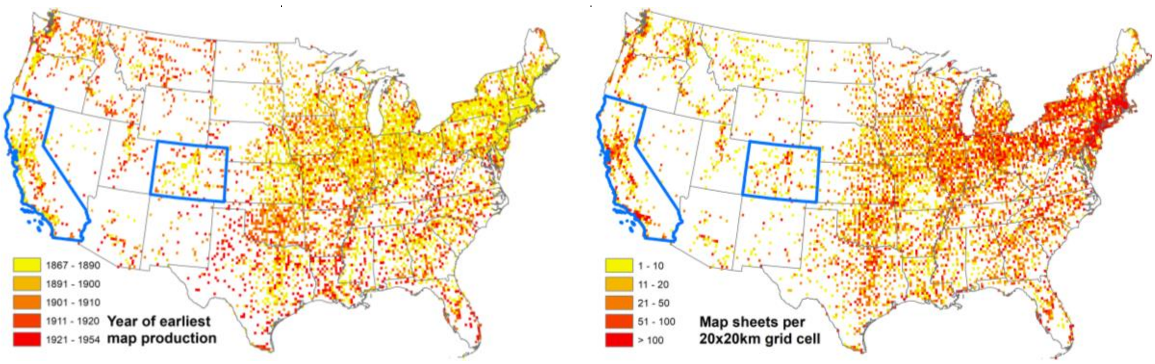


(a) (b) (c) (d)

Figure 7. (a) Map edition counts and **(b)** earliest map production year per 1:24,000 map quadrangle, **(c)** map edition counts and **(d)** earliest map production year per 1:62,500 map quadrangle in the state of California (USA) based on metadata analysis.

As a second example, the spatial-temporal coverage of the Sanborn fire insurance map archive is visualized. Since Sanborn map documents are commonly not offered as georeferenced datasets, this analysis is based on automatically extracted map locations (i.e., town or city name, county, and state) that were collected from unstructured metadata retrieved from HTML-based web content of the U.S. Library of Congress [45]. Additionally, the number of map sheets and their temporal coverage per location are extracted. The extracted data are geocoded using web-based geocoding services, which allows to visualize data availability and spatio-temporal coverage of Sanborn map documents. Figure 8 shows, similar to the above examples, the year of the first map production and the number of maps produced in total per location. The comparison of these visualizations for the highlighted states of California and Colorado to the previously shown Figures 6 and 7 shows the differences in spatio-temporal coverage between the two map archives, indicating a much sparser spatial coverage of the Sanborn map archive, but extending further back in time than the USGS map archive.

 doi:10.20944/preprints201803.0021.v1



(a) (b)

Figure 8. Sanborn fire insurance map archive coverage: **(a)** year of first map production per location and **(b)** number of available map sheets per location, both aggregated to grid cells of 10km for efficient visualization. Highlighted in blue the states of California and Colorado for comparison to the USGS map coverage shown in previous figures.

2.1.2. Metadata-based spatial-temporal analysis of positional accuracy

Positional accuracy of scanned maps can be caused by several factors, such as paper map distortions due to heat or humidity, the quality of surveying measurements on which the map production is based, deviations from the local geodetic datum at data acquisition time, cartographic generalization, and distortions introduced during the scanning and georeferencing process. While most of these effects cannot be reconstructed or quantified in detail, metadata delivered with the USGS topographic map archive contains information about the ground control points (GCPs) used for georeferencing the scanned map documents that allow for a partial assessment of these distortions and resulting positional inaccuracies.

The USGS topographic map quadrangle boundaries represent a graticule. For example, the corner coordinates for quadrangles of scale 1:24,000 are spaced in a regular grid of 7.5′x7.5′. Additionally, a finer graticule of 2.5′x2.5′ is depicted in the maps. The intersections of this fine graticule are used by the USGS to georeference the maps. Therefore, pixel coordinates at those locations (i.e., the GCPs) are collected, and the corresponding known world coordinates of the graticule intersections are used to establish a second-order polynomial transformation based on least-squares adjustment. This transformation is used to warp the scanned document into a georeferenced raster dataset. The GCP coordinate pairs are reported in the metadata, as well as an error estimate per GCP that provides information on the georeference accuracy in pixels. Based on these error estimates given in pixel units and the spatial resolution of the georeferenced raster given in meters, the root mean standard error (RMSE) reflecting georeference accuracy in meters is calculated. Appending these RMSE values as attributes to the map quadrangle polygons allows to visualize georeferenced accuracy across the spatial-temporal domain. This is shown for the USGS maps of scale 1:24,000 in the state of Colorado (Figure 9) for different time periods, visualizing the maximum RMSE per quadrangle and time period. Such temporally stratified representations are useful to examine if the georeference accuracy is constant over time. It can be seen that the earlier years in this example show higher degrees of inaccuracy than more recent map sheets. This has important implications for the user who is interested in using maps from different points in time that may exhibit different levels of inaccuracy.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 March 2018 doi:10.20944/preprints201803.0021.v1

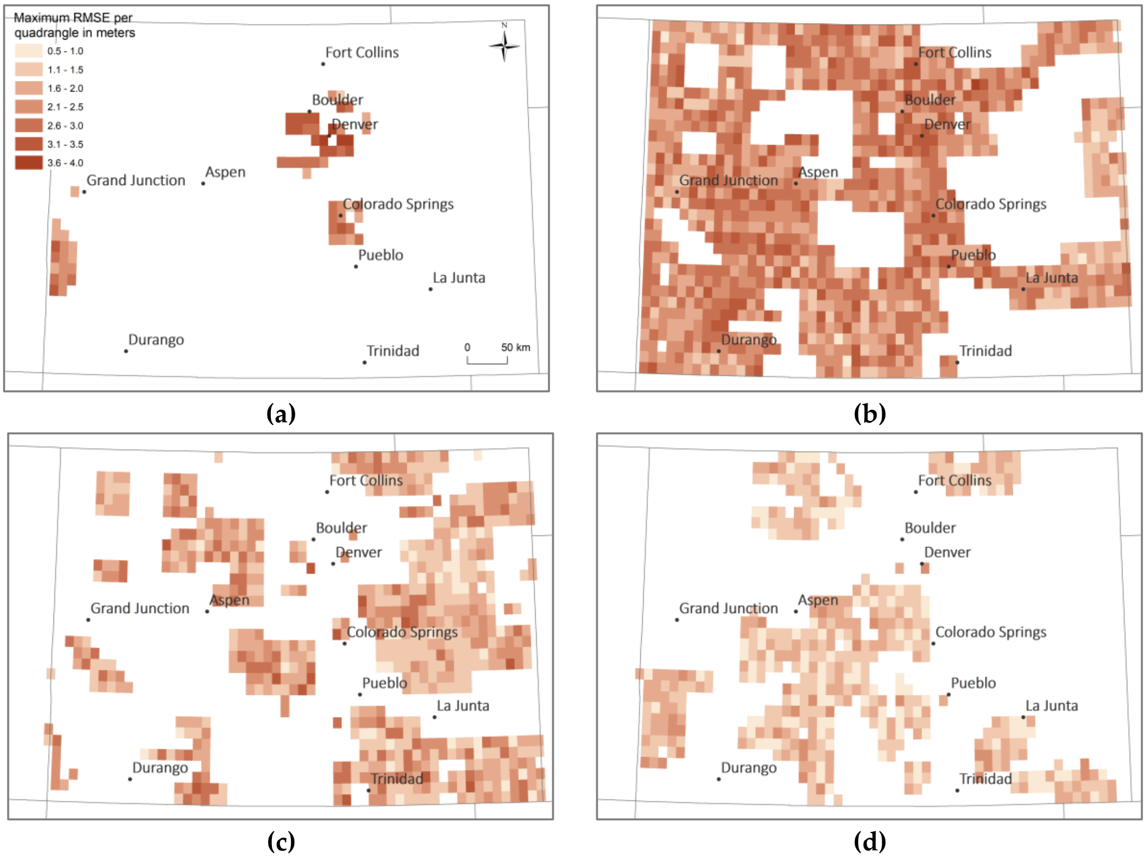


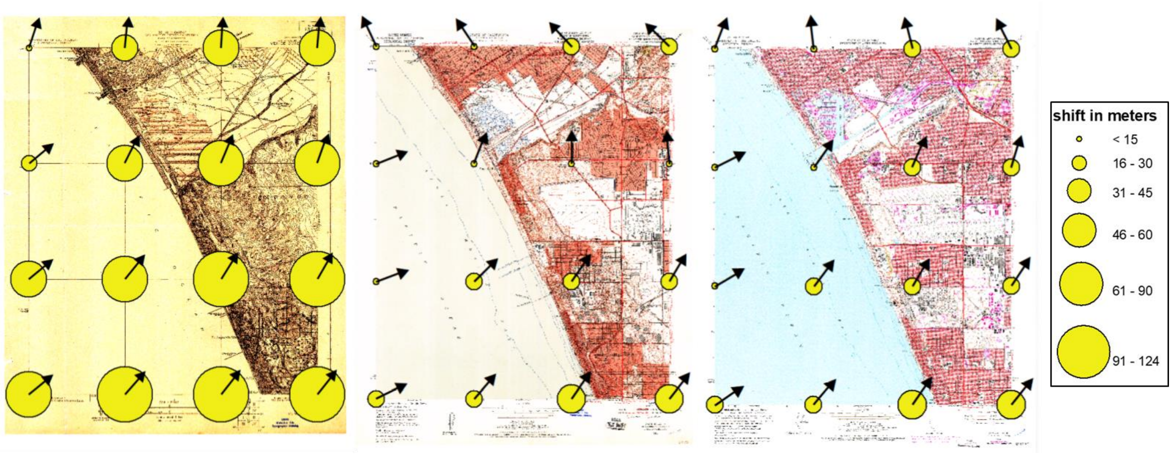
Figure 9. Spatio-temporal patterns of georeference accuracy of USGS topographic maps (1:24,000) in the state of Colorado (USA), for maps produced between **(a)** 1930-1950, **(b)** 1950-1970, **(c)** 1970-1990, and **(d)** 1990-2004.

Besides this information, the distortion introduced to the map by the warping process can be characterized by displacement vectors computed between the known world coordinates of each GCP (i.e., the graticule intersections) and the world coordinates corresponding to the respective pixel coordinates after applying the second-order polynomial transformation. These displacement vectors reflect geometric distortions and positional inaccuracy in the original map (i.e., *prior* to the georeferencing process) but are also affected by additional distortions introduced during georeferencing inaccuracies or through scanner miscalibration.

Assuming that objects in the map are affected by the same degree of inaccuracy like the graticule intersections, the magnitudes of these displacement vectors allow to estimate the maximum displacements to be expected between objects in the map and their real-world counterparts that may not be corrected by the second order polynomial transformation.

Figure 10 shows examples of these displacement vectors visualized for individual USGS map sheets at scale 1:24,000 from Venice (California) produced in 1923, 1957, and 1975. The magnitude of the local displacement is represented by the dot area, whereas the arrow indicates the displacement angle. This example shows similar patterns across the three maps, probably reflecting non-independent distortions between the maps since earlier maps are typically used as base maps for subsequent map editions, and some local variations due to inaccuracies introduced during georeferencing of the individual map sheets.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 March 2018 doi:10.20944/preprints201803.0021.v1



(a) (b) (c)

Figure 10. Displacement vectors at GCP locations characterizing the distortions introduced during the georeferencing of USGS topographic maps from Venice (California), produced in **(a)** 1923, **(b)** in 1957, and **(c)** in 1975 (from left to right).

Additionally, these displacement vectors can be visualized as vector fields across large areas, allowing to identify regions, quadrangles, or individual maps of high or low positional reliability, respectively. Figure 11 shows the vector field of relative displacements for USGS maps of scale 1:24,000 for a region Northwest of Denver, Colorado. Notable are the large displacement vectors in the Parshall quadrangle, indicating some anomalous map distortion, whereas the Cabin Creek quadrangle (Northeast of Parshall) seems to have suffered from very slight distortions only. Multiple arrows indicate the availability of multiple map editions in given quadrangles. Such visualizations may inform map users about the heterogeneity in distortions applied to the map sheets during the georeferencing process and may indicate different degrees of positional accuracy across a given study area.

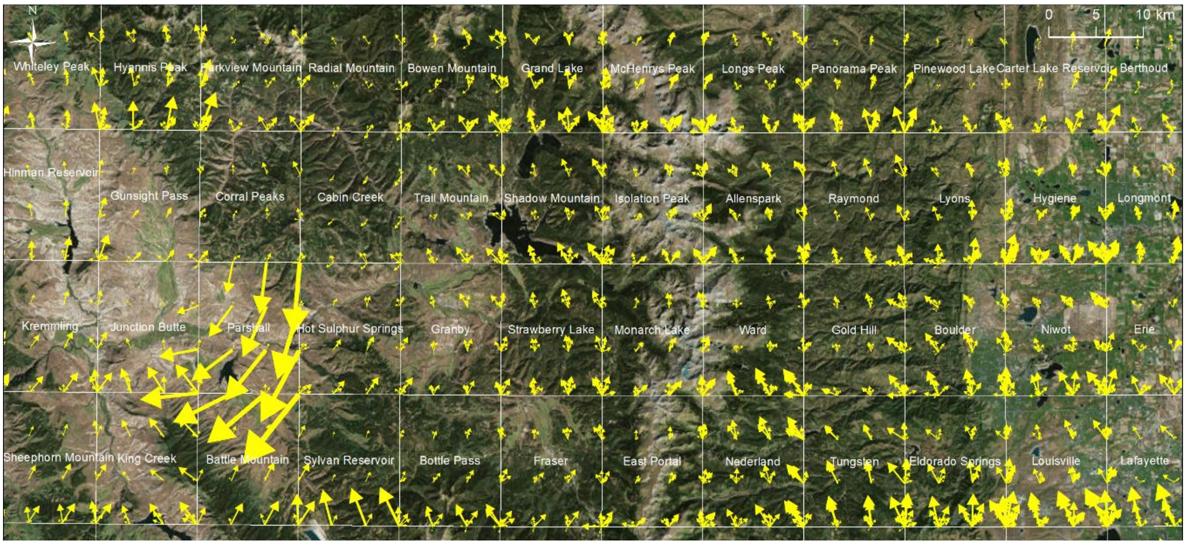


Figure 11. Displacement vector field at GCP locations over multiple USGS map quadrangles of scale 1:24,000, located North-west of Denver (Colorado), reflecting different types of distortions introduced to the map documents during the georeferencing process (Basemap source: [46]).

2.2. Content-based analysis

The presented metadata-based analysis provides valuable insights of spatial-temporal map availability, coverage, and spatial accuracy without analyzing the actual content of the map archives. However, it is important to inform the analyst about the degree of heterogeneity at content-level. In order to obtain detailed knowledge about the content of map archives, we propose a framework

 doi:10.20944/preprints201803.0021.v1

based on low-level image descriptors computed for each map or map patches. Here, we employ color-histogram based moments (i.e., mean, standard deviation, skewness and kurtosis, see [47]) computed for each image channel in the RGB color space. Mean and standard deviation characterize hue, brightness and contrast level of an image, skewness and kurtosis indicate the symmetry and flatness of the probability density of the color distributions, and thus reflect color spread and variability of an image. These four measures are computed for each channel of an image and stacked together to a 12-dimensional feature descriptor, at image or patch level. In the case of scanned map documents, such descriptors allow to retrieve maps or patches of maps of similar background color (depending on paper type and scan contrast level), and maps of similar dominant map content, such as waterbodies, urban areas, or forest cover. This similarity assessment is based on distances in the descriptor feature space and can involve metadata (e.g., map reference year), or ancillary geospatial data, to assess map content similarity across or within different geographic settings. Furthermore, approaches for dimensionality reduction such as t-distributed stochastic neighborhood embedding (t-SNE, [48]) are employed to visualize the data based on similarity in feature space. T-SNE allows to reduce the dimensionality of high-dimensional data, where the relative distances between the data points in the reduced feature space reflect the similarity of the data points in the original feature space. This facilitates the visual or quantitative identification of clusters of similar map sheets and provides a better understanding of the content of large map archives and their inherent variability. This kind of similarity assessment and metadata analysis is useful in generating knowledge which can be used to guide sampling designs to generate template or training data for supervised information extraction techniques.

2.2.1. Content-based analysis at map level

Analyzing the content of the entire map archive with respect to similarities between the individual map sheets is done by computing the image-moments based map descriptors. These 12-dimensional features are transformed into a reduced 3-dimensional feature space that can be visualized and interpreted intuitively. Figure 12 shows the 3D feature space for the 6,964 USGS maps in the state of Colorado. The map reference year is used to color-code the points representing individual map sheets. The clusters of dark blue points indicate fundamentally different color characteristics of old maps in comparison to more recent maps represented by points colored in green-yellow tones.

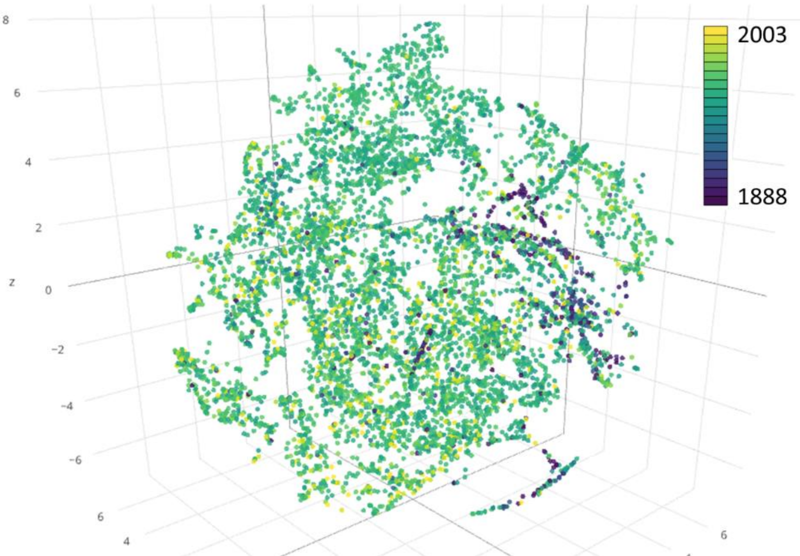


Figure 12. T-SNE visualization of the 6,964 USGS maps in the state of Colorado in a 3D feature space based on 12-dimensional image descriptors obtained from channel-wise image moments.

In addition to color-coding the data points by the corresponding map reference year, the 12-dimensional descriptors can be transformed into a 2D feature space and visualized using thumbnails

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 March 2018 doi:10.20944/preprints201803.0021.v1

of individual maps corresponding to each data point in Figure 12. This transformation results in an integrated visual assessment of map archives containing large numbers of map sheets. Figure 13 shows a t-SNE thumbnail visualization of a random sample (N=4,356) of the Colorado USGS maps in a 2D feature space. Nearest neighbor snapping is used to create a rectified visualization. This is a very effective way to visualize the variability in map contents, such as dominating forest area proportions. It also illustrates the presence and abundance of different map designs and base color use, e.g., high contrast and saturation levels in recent maps, compared to yellow-tainted map sheets from the beginning of the 20th century centered at the bottom. The latter corresponds to the cluster of historical maps located at the bottom of the point cloud in Figure 12.

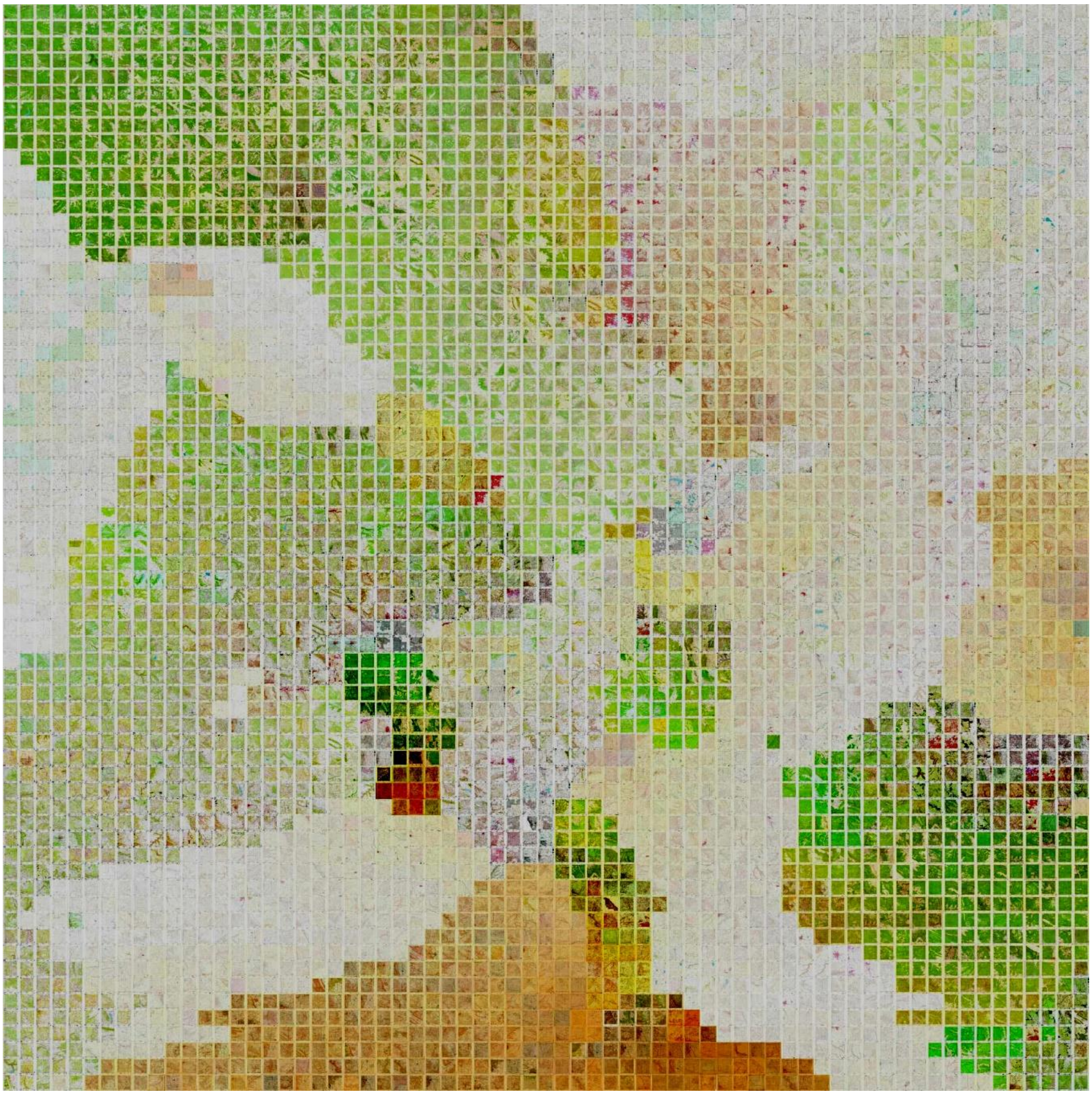


Figure 13. Thumbnail-based visualization of a subset of the USGS topographic maps in the state of Colorado (USA) based on a 2D transformation of the 12-dimensional image descriptor feature space using t-SNE.

2.2.2. Content-based analysis at within-map patch level

In order to assess the content within map sheets, map documents can be partitioned into tiles of a fixed size (exemplified here for 100x100 pixels). Low-level descriptors based on image moments can then be computed for each individual patch. However, if the patch size is chosen small enough, it is computationally feasible to use the raw (or down-sampled) patch data (e.g., a line vector of all pixel

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 March 2018 doi:10.20944/preprints201803.0021.v1

values in the patch) as a basis for t-SNE transformations. This can be useful if it is desired to introduce a higher degree of spatiality and even directionality when assessing the similarity between the patches. This method allows, for example, to rearrange a map document in patches based on patch similarity, as shown for an example USGS map in Figure 14a. Quadrangle boundaries based on corner coordinates delivered in the metadata can be used to clip the map contents and remove non-geographic content in the map sheet edges. The clipped map content is then partitioned in tiles down-sampled by factor 4, which results in a 1,875-dimensional feature vector per patch. These features are then transformed into a 2D-space using t-SNE in order to create a similarity-based rearrangement of the map patches (Figure 14b). This rearrangement highlights for example the groups of linear objects of different dominant directions, such as road objects, or clusters of patches that contain contour lines with diffuse directional characteristics. The incorporation of directionality may be useful to design sampling schemes that generate training data allowing for rotation-invariant feature learning.

(a) (b)

Figure 14. (a) USGS topographic map for Boulder, Colorado (1966), and **(b)** rearranged map patches according to their similarity in a raw pixel value feature space using t-SNE.

2.2.3. Content-based analysis at cross-map patch level

If variations of specific cartographic symbols across large map archives are of interest and have to be characterized, ancillary geospatial data can be employed to label the created map patches based on their spatial relationships to the ancillary data. For example, it may be important to assess the differences in cartographic representations of dense urban settlement areas across map sheets, in order to design a recognition model for urban settlement. In such situations building footprint data with built-year information and the respective spatio-temporal coverage can be employed to reconstruct settlement distributions in a given map reference year (see [49]). Based on these reference locations, building density surfaces can be computed for each map reference year. Using appropriate thresholding allows to approximately delineate dense settlement areas for a given point in time. Based on spatial overlap between map patches and these dense reference settlement areas, map patches that are likely to contain urban area symbols can be identified across multiple maps. These selected map patches can then be visualized in an integrated manner using t-SNE arrangement, as exemplarily shown in Figure 15 for map patches collected across 50 USGS maps (1:24,000) in the states of Colorado and California. This arrangement illustrates nicely the different cartographic styles that are used to represent dense urban settlements across time and map sheets, and provides valuable information useful for the design of a recognition model. Additional samples could be collected at locations where no ancillary data is available, and their content can be estimated based on descriptor similarity (i.e., patches of low Euclidean distance in the descriptor feature space) or using unsupervised or supervised classification methods.

Figure 15. T-SNE arrangement of cross-map samples of patches likely to contain dense urban settlement symbols.

3. Conclusions and Outlook

In this contribution we propose a set of methods for systematic information mining and content retrieval in large collections of cartographic documents, such as topographic map archives. These methods consist of pure metadata-based analyses, as well as content-based analyses using low-level image descriptors such as histogram-based color moments, and dimensionality reduction methods (i.e., t-SNE). We illustrate the proposed approach by exemplary analyses of the USGS topographic map archive and the Sanborn fire insurance map collection. Our approach can be used to explore and compare spatio-temporal coverage of these archives, the variability of positional accuracy, and differences in content of the map documents based on visual-analytical tools. These content-based map mining methods are inspired by image information mining systems implemented for remote sensing data archives and have been applied to facilitate the design and implementation of information extraction methods [27,28] Further work will include the identification of suitable textural measures to be incorporated in the image descriptors. Additionally, the benefit of map archive indexing based on low-level image descriptors will be tested in a prototypic map mining framework. Moreover, these efforts will contribute to the design of adequate sampling methods to generate representative training data for large-scale information extraction methods from historical map archives.

Acknowledgments: This material is based on research sponsored in part by the National Science Foundation under Grant Nos. IIS 1563933 (to the University of Colorado at Boulder) and IIS 1564164 (to the University of Southern California).

Author Contributions: J.H.U. and S.L. conceived and designed the experiments; J.H.U. performed the experiments; J.H.U. analyzed the data; J.H.U. wrote the paper.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Chiang, Y.-Y.; Leyk, S.; Knoblock, C.A. A survey of digital map processing techniques. *ACM Computing Surveys* **2014**, *47*, 1-44. http://dx.doi.org/10.1145/2557423
2. Miyoshi, T.; Weiqing, L.; Kaneda, K.; Yamashita, H.; Nakamae, E. Automatic extraction of buildings utilizing geometric features of a scanned topographic map. In *Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004.*, IEEE: 2004. http://dx.doi.org/10.1109/icpr.2004.1334607
3. Laycock, S.D.; Brown, P.G.; Laycock, R.G.; Day, A.M. Aligning archive maps and extracting footprints for analysis of historic urban environments. *Computers & Graphics* **2011**, *35*, 242-249. http://dx.doi.org/10.1016/j.cag.2011.01.002
4. Arteaga, M.G. Historical map polygon and feature extractor. In Proceedings of the 1st ACM SIGSPATIAL International Workshop on MapInteraction - MapInteract '13, ACM Press: 2013. http://dx.doi.org/10.1145/2534931.2534932
5. Chiang, Y.-Y.; Leyk, S.; Knoblock, C.A. Efficient and robust graphics recognition from historical maps. In *Graphics Recognition. New Trends and Challenges*, Springer Berlin Heidelberg: 2013; pp 25-35. http://dx.doi.org/10.1007/978-3-642-36824-0_3
6. Miao, Q.; Liu, T.; Song, J.; Gong, M.; Yang, Y. Guided superpixel method for topographic map processing. *IEEE Transactions on Geoscience and Remote Sensing* **2016**, *54*, 6265-6279. http://dx.doi.org/10.1109/tgrs.2016.2567481
7. Leyk, S.; Boesch, R. Extracting composite cartographic area features in low-quality maps. *Cartography and Geographic Information Science* **2009**, *36*, 71-79. http://dx.doi.org/10.1559/152304009787340115
8. Chiang, Y.-Y.; Moghaddam, S.; Gupta, S.; Fernandes, R.; Knoblock, C.A. From map images to geographic names. In *Proceedings of the 22nd ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems - SIGSPATIAL '14*, ACM Press: 2014. http://dx.doi.org/10.1145/2666310.2666374
9. Chiang, Y.-Y.; Leyk, S.; Honarvar Nazari, N.; Moghaddam, S.; Tan, T.X. Assessing the impact of graphical quality on automatic text recognition in digital maps. *Computers & Geosciences* **2016**, *93*, 21-35. http://dx.doi.org/10.1016/j.cageo.2016.04.013
10. Tsorlini, A.; Iosifescu, I. ; Iosifescu, C. ; Hurni, L. A methodological framework for analyzing digitally historical maps using data from different sources through an online interactive platform. *e-Perimetron*, **2014**, *9(4)*, 153-165.
11. Hurni, L. ; Lorenz, C. ; Oleggini, L. Cartographic reconstruction of historic settlement development by means of modern geo-data. Proceedings of the 26th International cartographic conference. Dresden, Germany, 2013.
12. Leyk, S.; and Chiang, Y. Information extraction of hydrographic features from historical map archives using the concept of geographic context. Proceedings of AutoCarto 2016, Albuquerque, New Mexico, USA, 2016.
13. Iosifescu, I.; Tsorlini, A. ; Hurni, L. Towards a comprehensive methodology for automatic vectorization of raster historical maps. *e-Perimetron* **2016**, *11(2)*, 57-76.
14. Budig, B.; van Dijk, T.C. Active learning for classifying template matches in historical maps. In *Discovery Science*, Springer International Publishing: 2015; pp 33-47. http://dx.doi.org/10.1007/978-3-319-24282-8_5
15. Budig, B.; Dijk, T.C.V.; Wolff, A. Matching labels and markers in historical maps. *ACM Transactions on Spatial Algorithms and Systems* **2016**, *2*, 1-24. http://dx.doi.org/10.1145/2994598
16. Budig, B.; van Dijk, T.C.; Feitsch, F.; Arteaga, M.G. Polygon consensus. In Proceedings of the 24th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems - GIS '16, ACM Press: 2016. http://dx.doi.org/10.1145/2996913.2996951
17. Maire, F.; Mejias, L.; Hodgson, A. A convolutional neural network for automatic analysis of aerial imagery. In *2014 International Conference on Digital Image Computing: Techniques and Applications (DICTA)*, IEEE: 2014. http://dx.doi.org/10.1109/dicta.2014.7008084
18. Castelluccio, M.; Poggi, G.; Sansone, C.; Verdoliva, L. Training convolutional neural networks for semantic classification of remote sensing imagery. In *2017 Joint Urban Remote Sensing Event (JURSE)*, IEEE: 2017. http://dx.doi.org/10.1109/jurse.2017.7924535
19. Audebert, N.; Le Saux, B.; Lefèvre, S. Semantic segmentation of earth observation data using multimodal and multi-scale deep networks. In *Computer Vision – ACCV 2016*, Springer International Publishing: 2017; pp 180-196. http://dx.doi.org/10.1007/978-3-319-54181-5_12

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 2 March 2018 doi:10.20944/preprints201803.0021.v1

20. Marmanis, D.; Datcu, M.; Esch, T.; Stilla, U. Deep learning earth observation classification using imagenet pretrained networks. *IEEE Geoscience and Remote Sensing Letters* **2016**, *13*, 105-109. http://dx.doi.org/10.1109/lgrs.2015.2499239
21. Romero, A.; Gatta, C.; Camps-Valls, G. Unsupervised deep feature extraction for remote sensing image classification. *IEEE Transactions on Geoscience and Remote Sensing* **2016**, *54*, 1349-1362. http://dx.doi.org/10.1109/tgrs.2015.2478379
22. Scott, G.J.; England, M.R.; Starms, W.A.; Marcum, R.A.; Davis, C.H. Training deep convolutional neural networks for land–cover classification of high-resolution imagery. *IEEE Geoscience and Remote Sensing Letters* **2017**, *14*, 549-553. http://dx.doi.org/10.1109/lgrs.2017.2657778
23. Zhao, W.; Jiao, L.; Ma, W.; Zhao, J.; Zhao, J.; Liu, H.; Cao, X.; Yang, S. Superpixel-based multiple local cnn for panchromatic and multispectral image classification. *IEEE Transactions on Geoscience and Remote Sensing* **2017**, *55*, 4141-4156. http://dx.doi.org/10.1109/tgrs.2017.2689018
24. Zhang, L.; Zhang, L.; Du, B. Deep learning for remote sensing data: A technical tutorial on the state of the art. *IEEE Geoscience and Remote Sensing Magazine* **2016**, *4*, 22-40. http://dx.doi.org/10.1109/mgrs.2016.2540798
25. Zhu, X.X.; Tuia, D.; Mou, L.; Xia, G.-S.; Zhang, L.; Xu, F.; Fraundorfer, F. Deep learning in remote sensing: A comprehensive review and list of resources. *IEEE Geoscience and Remote Sensing Magazine* **2017**, *5*, 8-36. http://dx.doi.org/10.1109/mgrs.2017.2762307
26. Ball, J.E.; Anderson, D.T.; Chan, C.S. Comprehensive survey of deep learning in remote sensing: Theories, tools, and challenges for the community. *Journal of Applied Remote Sensing* **2017**, *11*, 1. http://dx.doi.org/10.1117/1.jrs.11.042609
27. Uhl, J.H.; Leyk, S.; Yao-Yi, C.; Weiwei, D.; Knoblock, C.A. Extracting human settlement footprint from historical topographic map series using context-based machine learning. In *8th International Conference of Pattern Recognition Systems (ICPRS 2017)*, Institution of Engineering and Technology: 2017. http://dx.doi.org/10.1049/cp.2017.0144
28. Uhl, J.H.; Leyk, S.; Yao-Yi, C.; Weiwei, D.; Knoblock, C.A. Spatializing uncertainty in image segmentation using weakly supervised convolutional neural networks: a case study from historical map processing (under review)
29. Duan, W.; Chiang, Y.-Y.; Knoblock, C.A.; Jain, V.; Feldman, D.; Uhl, J.H.; Leyk, S. Automatic alignment of geographic features in contemporary vector data and historical maps. In *Proceedings of the 1st Workshop on Artificial Intelligence and Deep Learning for Geographic Knowledge Discovery - GeoAI '17*, ACM Press: 2017. http://dx.doi.org/10.1145/3149808.3149816
30. Duan, W.; Chiang, Y.-Y.; Knoblock, C.A.; Uhl, J.H.; Leyk, S. Automatic generation of precisely delineated geographic features from georeferenced historical maps using deep learning (under review)
31. Fishburn, K.A.; Davis, L.R.; Allord, G.J. Scanning and georeferencing historical usgs quadrangles. In *Fact Sheet*, US Geological Survey: 2017. http://dx.doi.org/10.3133/fs20173048
32. U.S. Library of Congress. Available online: http://www.loc.gov/rr/geogmap/sanborn/san6.html (accessed on 28/02/2018).
33. U.S. Library of Congress. Available online: http://www.loc.gov/rr/geogmap/sanborn/ (accessed on 28/02/2018).
34. U.S. Library of Congress. Available online: https://www.loc.gov/item/prn-17-074/sanborn-fire-insurance-maps-now-online/2017-05-25/ accessed on 28/02/2018).
35. National Library of Scotland. Available online: https://maps.nls.uk/os/index.html (accessed on 28/02/2018).
36. National Library of Scotland. Available online: http://maps.nls.uk/geo/explore (accessed on 28/02/2018).
37. Datcu, M.; Daschiel, H.; Pelizzari, A.; Quartulli, M.; Galoppo, A.; Colapicchioni, A.; Pastori, M.; Seidel, K.; Marchetti, P.G.; D'Elia, S. Information mining in remote sensing image archives: System concepts. *IEEE Transactions on Geoscience and Remote Sensing* **2003**, *41*, 2923-2936. http://dx.doi.org/10.1109/tgrs.2003.817197
38. Quartulli, M.; G. Olaizola, I. A review of eo image information mining. *ISPRS Journal of Photogrammetry and Remote Sensing* **2013**, *75*, 11-28. http://dx.doi.org/10.1016/j.isprsjprs.2012.09.010
39. Espinoza-Molina, D.; Alonso, K.; Datcu, M. Visual data mining for feature space exploration using in-situ data. In *2016 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, IEEE: 2016. http://dx.doi.org/10.1109/igarss.2016.7730543
40. Espinoza Molina, D.; Datcu, M. Data mining and knowledge discovery tools for exploiting big earth-observation data. *ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences* **2015**, *XL-7/W3*, 627-633. http://dx.doi.org/10.5194/isprsarchives-xl-7-w3-627-2015

 doi:10.20944/preprints201803.0021.v1

41. Griparis, A.; Faur, D.; Datcu, M. Dimensionality reduction for visual data mining of earth observation archives. *IEEE Geoscience and Remote Sensing Letters* **2016**, *13*, 1701-1705. http://dx.doi.org/10.1109/lgrs.2016.2604919
42. Durbha, S.S.; King, R.L. Semantics-enabled framework for knowledge discovery from earth observation data archives. *IEEE Transactions on Geoscience and Remote Sensing* **2005**, *43*, 2563-2572. http://dx.doi.org/10.1109/tgrs.2005.847908
43. Kurte, K.R.; Durbha, S.S.; King, R.L.; Younan, N.H.; Vatsavai, R. Semantics-enabled framework for spatial image information mining of linked earth observation data. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* **2017**, *10*, 29-44. http://dx.doi.org/10.1109/jstars.2016.2547992
44. Silva, M.P.S.; Camara, G.; Souza, R.C.M.; Valeriano, D.M.; Escada, M.I.S. Mining patterns of change in remote sensing image databases. In *Fifth IEEE International Conference on Data Mining (ICDM'05)*, IEEE. http://dx.doi.org/10.1109/icdm.2005.98
45. Library of Congress. Available online: http://www.loc.gov/rr/geogmap/sanborn/country.php?countryID=1 (accessed on 28/02/2018).
46. ESRI Basemaps: Esri, DigitalGlobe, GeoEye, Earthstar Geographics, CNES/Airbus DS, USDA, USGS, AEX, Getmapping, Aerogrid, IGN, IGP, swisstopo, and the GIS User Community
47. Huang, Z.-C.; Chan, P.P.K.; Ng, W.W.Y.; Yeung, D.S. Content-based image retrieval using color moment and gabor texture feature. In *2010 International Conference on Machine Learning and Cybernetics*, IEEE: 2010. http://dx.doi.org/10.1109/icmlc.2010.5580566
48. Van der Maaten, L; Hinton, G. Visualizing data using t-SNE. *Journal of Machine Learning Research*, **2008**, 9, 2579-2605
49. Leyk, S.; Uhl, J.H.; Balk, D.; Jones, B. Assessing the accuracy of multi-temporal built-up land layers across rural-urban trajectories in the united states. *Remote Sensing of Environment* **2018**, *204*, 898-917. http://dx.doi.org/10.1016/j.rse.2017.08.035