International Workshop on Automatic Vectorisation of Historical Maps - 13 March 2020 - ELTE, Budapest
https://doi.org/10.21862/avhm2020.01

Drolias Garyfallos Chrysovalantis[1], Tziokas Nikolaos[2]

Building footprint extraction from historic maps utilizing automatic vectorisation methods in open source GIS software

Keywords: automatic vectorisation; historical maps; open source GIS; shape recognition

Summary: Historical maps are a great source of information about city landscapes and the way it changes through time. These analog maps contain rich cartographic information but not in a format allowing analysis and interpretation. A common technique to access this information is the manual digitization of the scanned map but it is a time-consuming process. Automatic digitization/vectorisation processing generally require expert knowledge in order to fine-tune parameters of the applied shapes recognition techniques and thus are not readily usable for non-expert users. The methodology proposed in this paper offers fast and automated conversion from a scanned image (raster format) to geospatial datasets (vector format) requiring minimal time for building footprint extraction. The vector data extracted quite accurate building footprints from the scanned map, but the need for further post-processing for optimal results, is imminent.

Introduction

Historical maps contain a lot of information about a region in a certain period, depict the condition of a settlement and toponyms of the time the map was created. Not only those maps are a part of our cultural heritage, but some of them are of great interest to researchers from various academic fields. Maps record the geographical information that is essential to reconstruct past places, town-wide, region-wide or even nation-wide depending on the map's scale and extent. These maps often contain information retained by no other written source, such as toponyms, boundaries, and geomorphological features that have undergone any type of change or even elimination through time. A map's degree of accuracy is a genuine source of information about the state of technology and scientific understanding at the time of its creation. By incorporating information from historical maps and open-source GIS software, researchers are now able to extract a vast amount of information, analyze and interpret with other geospatial data.

Most of those maps are found in analogue form, thus they need to be processed in order to extract geospatial data for further analysis and interpretation. Scanning of historical maps is usually the first step towards converting them in a format allowing analysis in GIS platforms. When a map is scanned, a raster image is created. Raster images are a matrix of pixels with each pixel having a data value. A raster image can be black-and-white, containing a single band, or colored containing multiple bands (red, green, blue). A black-and-white image often is a binary image, with each pixel having the value of 1 or 0, but a multicolor image means that each pixel has a data value for each band. Conversely, vector data are in the form of points, lines, and areas (polygons). Points represent discrete locations, lines represent discrete linear features, and polygons represent discrete bounded areas. Raster images resulting from scanning methods need to be converted in vector format in order to extract further information from the historic map.

[1] University of the Aegean [akisd@hotmail.com]
[2] University of the Aegean [nikos.tziokas@gmail.com]

https://doi.org/10.21862/avhm2020.01

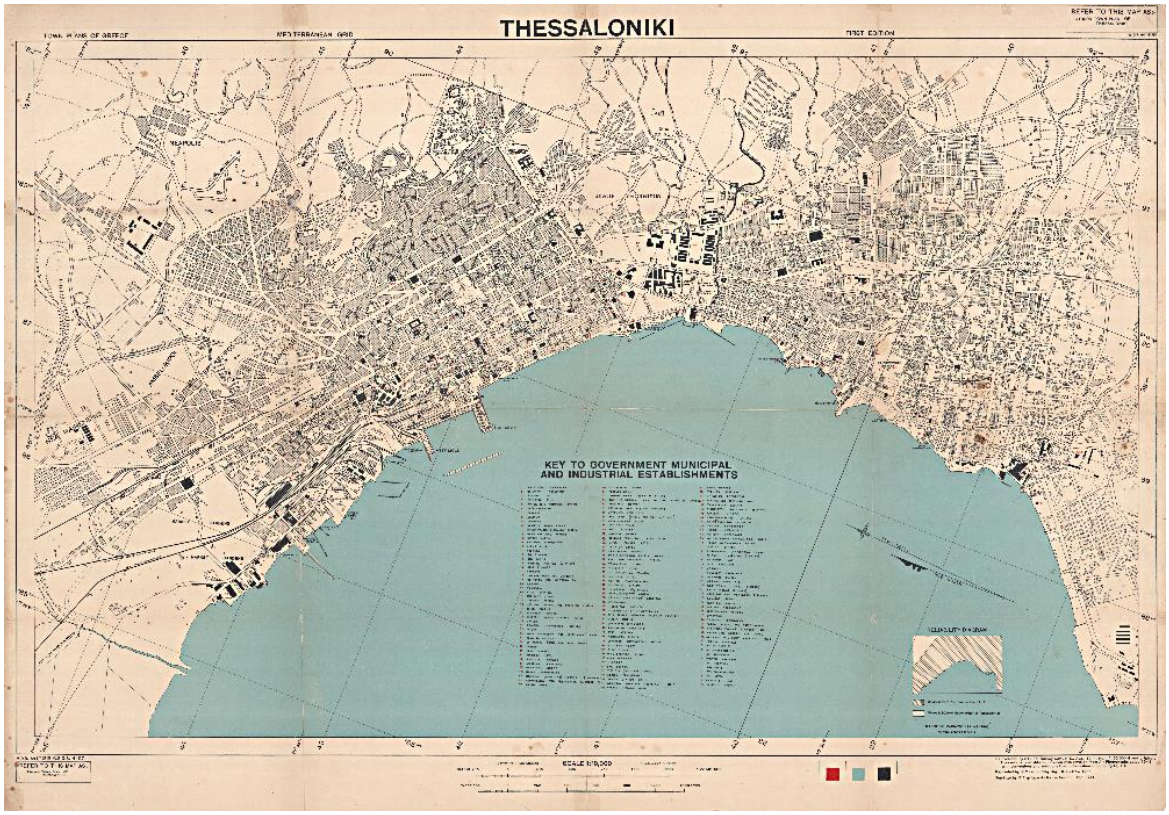

Figure 1. Original scanned map

This digitization method usually is the most challenging part as it can often lead to poor results, depending on map quality, conversion types and methodologies applied to the initiative raster image.

In this paper, a new methodology is proposed based on free and open source GIS platforms in order to convert scanned (raster images) historical maps to manageable geospatial data (vector format). The basic stages of this procedure are the scanning and the correct georeference of the map, the pre-processing and the removal of pixel clumps in the raster image so that building foot-prints are extracted more efficiently. The second part of our methodology focuses on correcting the geometry of the converted vector data. Finally, a model containing all the processing steps and algorithms is developed in an open source GIS platform. This methodology has been applied on a single historic map, although it could be effective in any map containing information of a settlement, where buildings are represented as polygons.

Chiang et al., (2013) sought to automatically extract the road network from a USGS historical topographic map. They used the Color Image Segmentation (CIS) technique, which belongs to the shape recognition family, where the technique separates homogeneous color areas of the map into groups. Godfrey & Eveleth (2015) used Image enhancement techniques in a GIS environment to make regions of similar color more consistent, and then performed an unsupervised classification to group (pixels) at a similar value based on their color. Kim et al., (2014) extracted parcels from old cadastral maps by segmentation. First, they removed the grid reference lines to dilute the image, then the labels considering the morphological and geometric features of the diluted image, and then rebuilt the boundaries of the land while the terrestrial areas of interest of a user were shaped by their polygonal approaches. Finally, Iosifescu et al., (2016) extracted information from various features of the Topographic Map of the Canton of Zurich. Summarizing their proposed methodology: scanning and geo-referencing the map, use of morphology of shapes to identify lines and polygons, vector creation and elimination of outliers and spikes.

https://doi.org/10.21862/avhm2020.01

Data

In our project, we used an already scanned map provided by The Cartographic Heritage Archives in order to apply our proposed methodology for the extraction of building footprints from a historical map. It is a town plan of Thessaloniki in scale 1:10.000, reproduced and reprinted by Great Britain Army Royal Engineers Fd. Survey Coy., 512 in 1944, created with revision from air photographs dated 1943. This is a multicolor map focusing on government municipal and industrial establishments of Thessaloniki, represented as single color filled polygons, while building blocks are represented as pattern-filled polygons. Each building is tagged with a number corresponding to its description in the map legend.

Methodology

Overview

Automatic vectorisation of raster images is not a simple task, considering the quality of the image and the way features such as buildings are depicted on it. In our case the raster image is already scanned in 300 dots per inch, satisfying enough to proceed with our proposed methodology. Our methodology consists of four fundamental steps (Figure2). Each step is carefully chosen in order to deliver the desired results. Constantly validating the intermediate results from each step and fine-tuning the parameters of each toolbox implemented, is fundamental in order to create a final model able to extract vector data from a raster image. In QGIS model builder, a variety of tools from GRASS and SAGA can be imported in the same model, allowing us to run all of the processing stages as one single process. These stages are:

1. Georeference of the scanned map; 2. Raster image processing; 3. Conversion from raster to vector format; 4. Vector data processing.

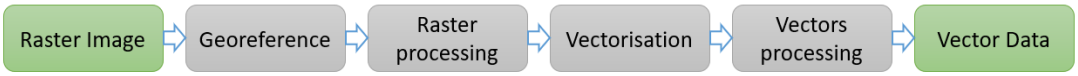

Figure 2. Methodology overview

Georeference

Integrating historical maps in GIS to analyze the spatial information they contain, or to incorporate them with other spatial data, requires georeferencing. That is, selected control points on the scanned image must be aligned with their actual geographical location, either by assigning geographical coordinates to each point, or by linking each point to its equivalent on a digital map. Once the control points are in place, mathematical algorithms are applied to warp the original raster image to fit the chosen map projection as nearly as possible. Further adjustments can be done manually to try to find the best fit for all parts of the original map.

In our case a total of 11 control points were selected distributed equally throughout the scanned image. Corresponding coordinates were selected manually using a basemap developed from Hellenic Cadaster with a pixel size of approximately 50 cm, and Greek Geodetic Refe-rence System 1987 (GGRS87 / Greek Grid) as our reference system. Map units in this system are measured in meters. Choosing carefully places that exist both in the basemap and the his-torical map and moving control points to their corresponding coordinates resulting in mean error 0,894282 map units or 1,05941 pixels, in order to apply second-order polynomial trans-formation and nearest neighbor sampling method in QGIS Georeferencer toolbox. The output from this stage is a georeferenced tagged image file format (TIFF) raster image with no compression.

https://doi.org/10.21862/avhm2020.01

Table 1. Control points coordinates and residuals

ID	Enabled	Pixel X	Pixel Y	Map X	Map Y	Res X (m)	Res Y (m)	Res Total (m)
0	yes	5728	-3032	410905.088	4497529.053	0.16047	-0.0616369	0.171901
1	yes	10312	-6466	409839.129	4492794.061	0.151569	0.497035	0.519631
2	yes	4541	-1192	411916.473	4499101.103	0.331448	0.324183	0.463629
3	yes	4207	-774	412123.748	4499508.069	0.200376	-0.0958679	0.222128
4	yes	4841	-1555	411739.698	4498741.246	-0.0259153	-0.10182	0.105066
5	yes	3788	-2558	410598.486	4499176.636	-0.399351	-0.900828	0.985379
6	yes	7151	-2020	412182.433	4496815.487	-0.559665	0.345404	0.657669
7	yes	7184	-2564	411772.514	4496593.600	0.0757101	-0.380722	0.388177
8	yes	3161	-3366	409755.892	4499357.962	-0.864928	0.429498	0.965696
9	yes	2170	-2417	410144.813	4500455.658	0.514763	1.17594	1.28367
10	yes	4533	-3969	409764.555	4498097.663	0.415524	-1.23118	1.29941

Raster image processing

The next step of the image processing in order to apply our vectorisation method is the reclassifi-cation of the raster image and the conversion to binary image so as to extract building footprints from the original map. This is a significant first step towards extracting valuable information from our historic map as it separates the pixels containing information from the background of the map.

Reclassification of the raster image is performed applying rules, setting up a threshold initially in each band (red, green, blue) targeting to classify the image into two classes, one containing the buildings and one for every other value. Inquiring some pixels within color (black) filled polygons, in our map depicting buildings, we can identify that the maximum value a pixel can have is 80 out of the maximum of 255 in RGB scale. Since this value is the same at every band in black color there is no need to process the image separating each band or using any other color image segmentation technique. This is why we use the r.reclass toolbox from GRASS GIS, setting the following reclass rules: 0-80=1 , 81-255=0.

The reclassified raster image can then be negated using r.null toolbox from GRASS GIS in order to convert the class containing pixels with value equal to zero, as no-data. This step is greatly decreasing the size of the image, while ensuring our further applied processing will be applied in parts of the image containing pixels with information and not background values.

The binary image resulted from the previous steps needs some cleaning before it is ready for the vectorisation process. Although many filters can be applied, “remove small pixel clumps to no-data” toolbox from GRASS GIS seems to deliver the best results in our case. Removing groups of scattered pixels while not erasing a single pixel representing a building or generally part of a polygon. Thresholds here could be set from 50 to 350 regarding how many pixels represent a building. All groups of pixels with a total count of pixels below this threshold are set to no-data, removing many unwanted labels, letters in the map or even pixels scattered outside polygons. In this case we set the threshold at 200 considering optimal results without erasing a single building throughout the map.

https://doi.org/10.21862/avhm2020.01

Vector data processing

In this step a simplification procedure was performed in order to create more smooth lines for the features. The most commonly used algorithm in cartography is the Douglas-Peucker (DP) (Shi & Cheung, 2006). The algorithm splits the line data recursively and controls the compression quality by the threshold, and it is widely used in simplifying the trajectory of moving point objects due to its speed and accuracy. The DP algorithm constitutes of the following steps:

1. The first and the last point of the polyline are considered fixed and are linked by a straight line. The distances between the remaining points and the straight line is computed.
2. A tolerance distance is defined by the user and the distance of the furthest point is examined; if this distance is greater than a predefined tolerance, the point is added as a new vertex of the output polyline.
3. The iteration is now complete and the algorithm starts from the step 1, except that the new points created from the step 2 are now considered fixed. If no other points exceed the tolerance distance the algorithm concludes.

The last step was the filling of the holes (i.e. building block) less than a predefined size.

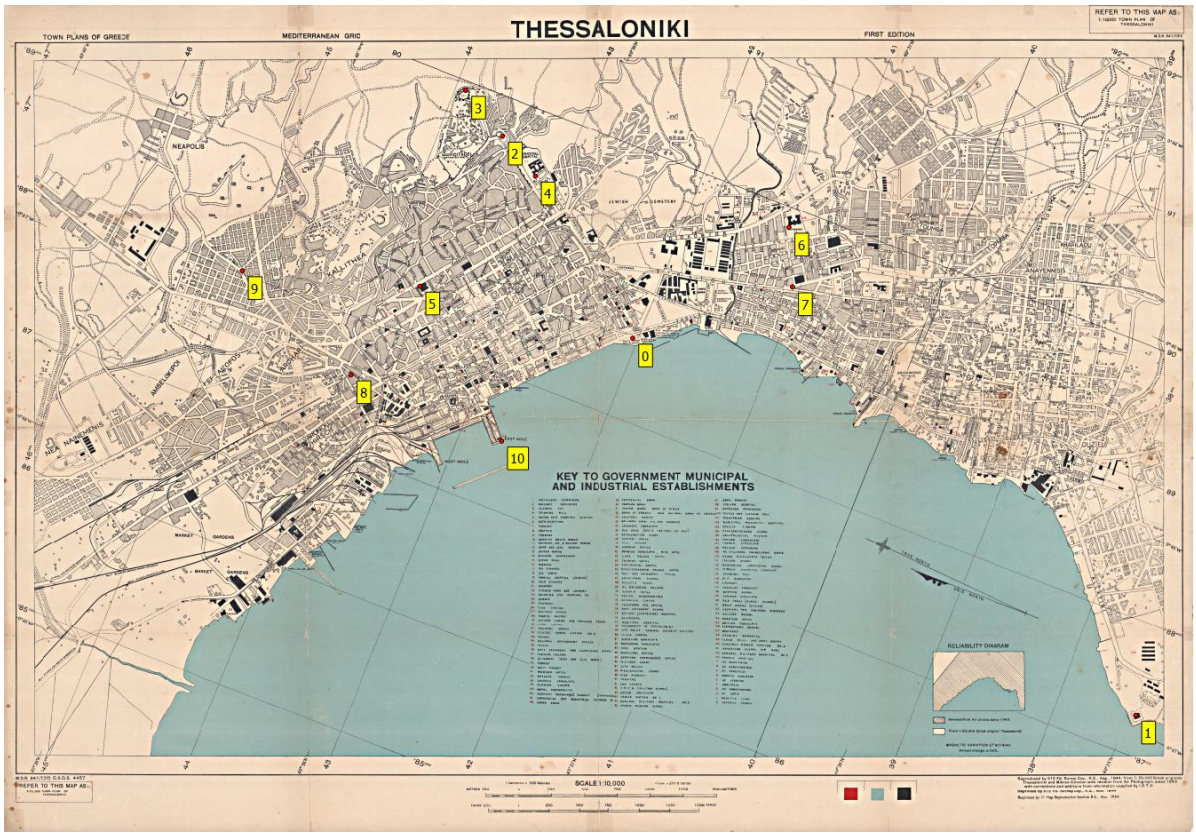

Figure 3. Control points used in Georeference

https://doi.org/10.21862/avhm2020.01

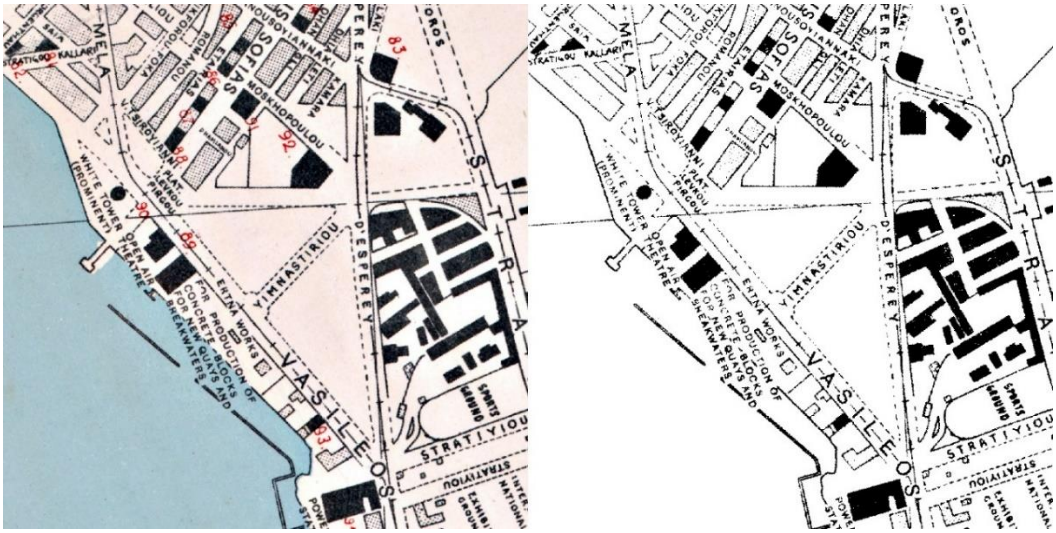

Figure 4. Left – scanned RGB image, Right – binary image

Figure 5. Left – binary image, Center – clumps marked as red, Right – binary cleaned image

https://doi.org/10.21862/avhm2020.01

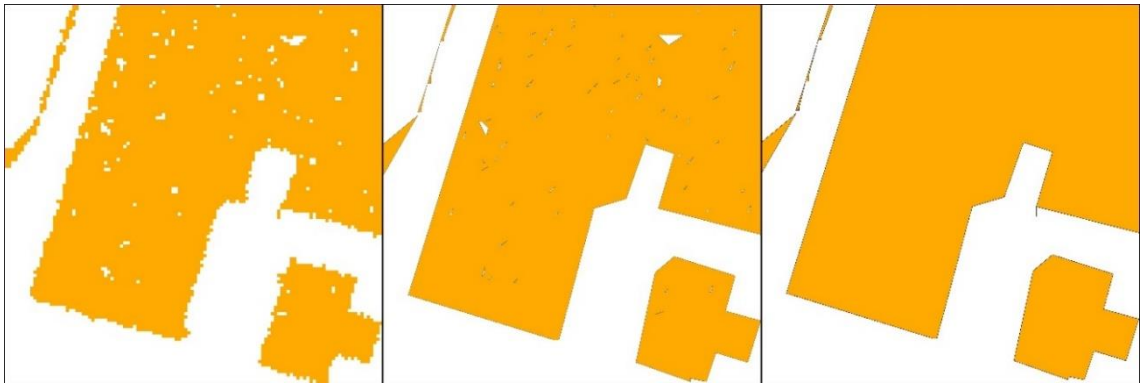

Figure 6. Left – Vectors extracted from cleaned binary image, Center – Simplified polygon lines, Right – Final results after the filling of holes

Model in QGIS

Final step of our proposed methodology is to implement all the processing described above into one model in QGIS. After running each step separately in order to fine-tune parameters and evaluate intermediate results, now we can use graphical modeler in QGIS to incorporate a series of tools to further automatize the vectorisation process meeting our needs for building footprints extraction from a scanned historic map.

In this case we insert all the tools from QGIS, GRASS GIS and SAGA GIS we used in the previous steps with the parameters described in each step. The only difference is the input of one more r.reclass and r.null tools in order to convert the filtered raster after the removal from clumps to no-data into a binary image.

Testing our model in a personal computer with eight-core processor, 16 gigabytes of RAM, and SSD disk, the results are more than promising. The georeferenced raster image of 700 megabytes, used as input, after a processing time of 2 minutes and 48 seconds, extracted vector data depicting building footprints quite accurately.

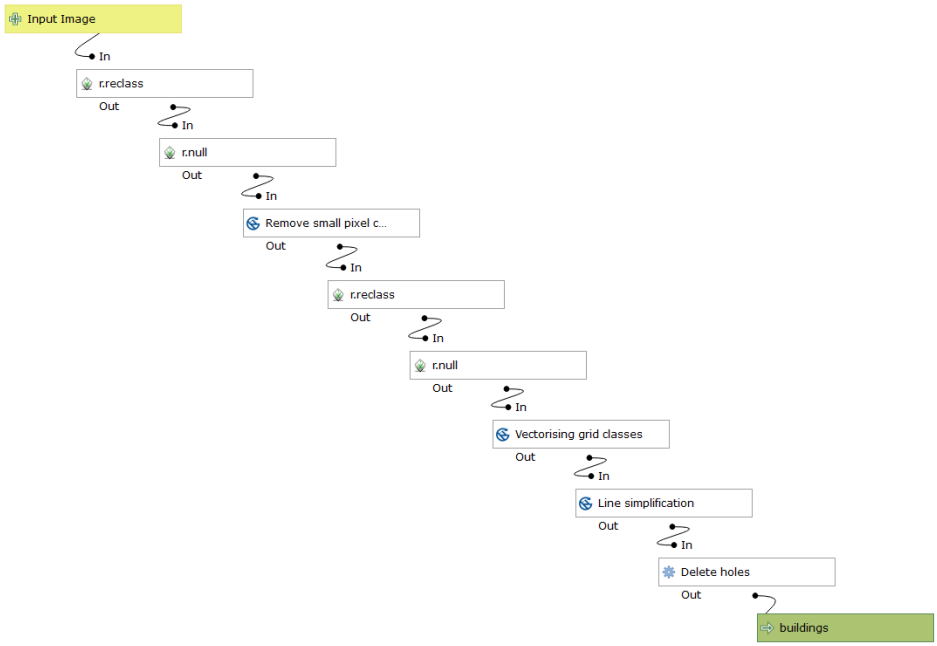

Figure 7. Model in QGIS Graphical Modeler

https://doi.org/10.21862/avhm2020.01

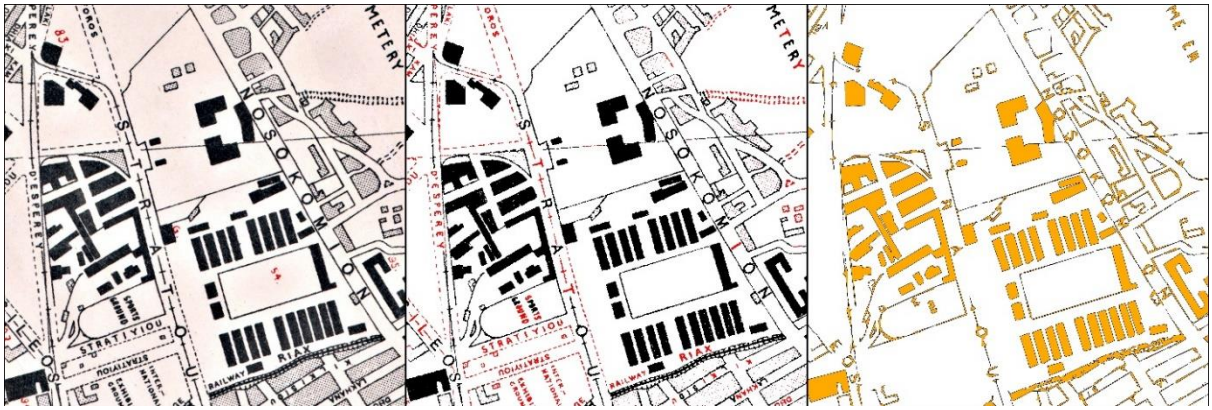

Figure 8. Left – Scanned image, Center – Clumps removed in red color, Right – Vectors extracted

Conclusions

In this paper we have presented an applied, semiautomated, practical workflow for generating polygon vector features from raster images implementing tools and algorithms from GRASS GIS and SAGA GIS into graphical modeler in QGIS. This general workflow can be used by researchers in generating vector data sets from thematic map raster images in hopes of unlocking some of the information contained in historical maps. Users need to modify parameters of the individual geoprocessing steps within the model to achieve the desirable result on a different map based on RGB values of the scanned image, but the general approach should prove valuable. QGIS is an open source GIS platform, thus allowing a wide variety of users to perform geoprocessing functions in several applications. Therefore, the threshold to modify this workflow, albeit adjusting the parameters to fit specific values, should be attainable for those interested in extracting valuable information from a scanned historical map within minimum amount of time. Based on the individual needs for automatic vectorisation and the available data, further tools could be added in the proposed model.

In some parts of the map, gridlines from the scanned image, text connected with polygons or strikethrough text presenting toponyms and other labels could not be removed with our methodology. In order to completely clear the image of any artifact not presenting a building polygon, manual selection of these polygons could help remove them. Text can not be automatically removed, so further processing is needed in order to totally remove any letter in the scanned image.

Acknowledgment

Authors would like to thank The Cartographic Heritage Archives, for providing an already scanned historical map ready to apply our proposed methodology

References

Ramer, U. (1972): An iterative procedure for the polygonal approximation of plane curves. Computer Graphics and Image Processing, 1(3), 244-256.
https://doi.org/10.1016/S0146-664X(72)80017-0

Douglas, D., Peucker, T. (1973): Algorithms for the reduction of the number of points required to represent a digitized line or its caricature. The Canadian Cartographer 10(2), 112-122. https://doi.org/10.3138/FM57-6770-U75U-7727

Yao-Yi, C., Leyk, S., and Knoblock, C.A., "A Survey of Digital Map Processing Techniques." *ACM Computing Surveys* 47, no. 1 (May 1, 2014): 1–44. https://doi.org/10.1145/2557423

Kim, Nam Wook, Jeongjin Lee, Hyungmin Lee, and Jinwook Seo. "Accurate Segmentation of Land Regions in Historical Cadastral Maps." *Journal of Visual Communication and Image Representation* 25, no. 5 (July 2014): 1262–74. https://doi.org/10.1016/j.jvcir.2014.01.001

Godfrey, B., and Eveleth, H., "An Adaptable Approach for Generating Vector Features from Scanned Historical Thematic Maps Using Image Enhancement and Remote Sensing Techniques in a Geographic Information System." *Journal of Map & Geography Libraries* 11, no. 1 (January 2, 2015): 18–36. https://doi.org/10.1080/15420353.2014.1001107

Conrad, O., B. Bechtel, M. Bock, H. Dietrich, E. Fischer, L. Gerlitz, J. Wehberg, V. Wichmann, and J. Böhner. "System for Automated Geoscientific Analyses (SAGA) v. 2.1.4." *Geoscientific Model Development* 8, no. 7 (July 7, 2015): 1991–2007. https://doi.org/10.5194/gmd-8-1991-2015

Iosifescu, I., Tsorlini, A., and Hurni, L., "Towards a Comprehensive Methodology for Automatic Vectorization of Raster Historical Maps." *EPerimetron* 11 (2016): 20.

Pallero, J.L.G. "Robust Line Simplification on the Plane." *Computers & Geosciences* 61 (December 2013): 152–59. https://doi.org/10.1016/j.cageo.2013.08.011

Shi, W., and Cheung, C.K., "Performance Evaluation of Line Simplification Algorithms for Vector Generalization." *The Cartographic Journal* 43, no. 1 (March 2006): 27–44. https://doi.org/10.1179/000870406X93490