AUTOMATIC ROAD EXTRACTION AND VECTORIZATION FROM SCANNED TOPOGRAPHIC MAPS

Marko Mrlješ[1]*,
Miloš Basarić[1],
Saša Bakrač[1],
Stevan Radojčić[1, 2]

[1]Military Geographical Institute - "General Stevan Bošković", Belgrade, Serbia

[2]Academy of Technical and Art Applied Studies, School of Civil Engineering and Geodesy, Belgrade, Serbia

Abstract:

The previous practice has shown that older editions of topographic maps of the Military Geographical Institute can be used as a reliable source of data in the process of creating digital topographic maps at a scale of 1:25 000. Since vectorization of the topographic data for the complete scale series of topographic maps has not been done, these maps are used exclusively in the form of raster services as a supplement to primary cartographic sources. Therefore, the large amount of data on these topographic maps can provide insight into how some phenomena moved and developed over time, but remained in raster form unsuitable for processing, analysis, and comparison with vector data obtained from primary cartographic sources and depicting the real character of some occurrences. The paper shows the process by which, using the Python programming language in combination with the ArcScan vectorization tool, line symbols of paved roads are extracted from the 1:25 000 topographic maps and then translated into a vector form suitable for further use. The data obtained in this way become available for multiple applications with a great saving of time, considering that the process is completely automated. A proposal for using the results through a comparison of the higher-order road network between the situation in 1969 and the situation in 2022 is given.

Keywords:

Object Detection, Automatic Vectorization, Scanned Maps Processing, Computer Vision, Image Processing.

INTRODUCTION

MGI is a scientific institution with a tradition of more than 140 years, whose current main activity is the production of topographic maps on a scale of 1:25 000. This map records 4 editions, so the I and II editions were made using the technological methods of the time in the period from 1947 to 1980 [1], while after 2002 it switched to the production of digital topographic maps (part III and IV edition). The transition from analog to digital production process happened with the III edition. The production process of the IV edition of the map (hereinafter DTM25) was established in the period from 2012 to 2015 [2] is still ongoing, and although the initially established digital techniques have proven themselves well in their beginnings, there is a justified need to improve the process in order to achieve faster and higher production as concluded in [3].

Correspondence:
Marko Mrlješ

e-mail:
markodmrljes@gmail.com

DOI: 10.15308/Sinteza-2023-227-234

The aim is to implement digital image processing and object detection methods, machine learning methods, and advanced cartographic processing methods to upgrade the existing production process. Despite the progress of technological processes, the II edition of the topographic map 1:25 000 (hereinafter TM25), due to the rich content and accuracy of the data, still finds its application in the current production process as the main secondary source data. 2051 sheets were produced in this edition and it covers the entire territory of the former Yugoslavia [1]. In the period from 2002 to 2004, the translation of TM25 into digital form was started, by scanning map sheets. After scanning, georeferencing was performed in the national coordinate system within the seventh zone of the Gauss-Kriger projection (creation of *.tif and *.tfw files), and later in the UTM coordinate system [4].

The application of TM25 for the production of DTM25 can be viewed in two directions. The first one takes into account the obsolescence of TM25 and the differences in the content that are evident and uses them as a basis for an analytical-comparative overview of the changes in phenomena and infrastructure in relation to the state of content on DTM25, which is up-to-date. This argument is grounded in the idea that without additional processing, it is impossible to perform quantitative and geometrical analysis based on these maps [5]. This direction stems from the fact that TM25 is used in the form of a raster service, while the creation of vector content based on it for the entire area it covers was absent due to the orientation of human resources to the production of new topographic content for DTM25, rather than the vectorization of the TM25. Therefore, the analyzes in this manner were limited to small parts of the territory. The second concerns the application of methods of computer detection of objects and details on TK25 in raster form and obtaining vector spatial data that can be the starting point for creating DTM25 during 2D restitution. It must be borne in mind that the second approach is favorable for infrastructural contents that are less subject to changes such as facilities and communications. Both directions are based on some kind of detection from raster forms and the application of advanced methods of automatic vectorization.

While the application of machine learning methods for multiclass segmentation of raster maps [6, 7] is popular and practiced, this paper used a simpler color-based segmentation method known as the CIS method [7]. This paper will propose a methodology for the automatic vectorization of higher-order roads on TM25 sheets, representing the wider region of the city of Kruševac. The focus is on roads that, according to their categorization, have great infrastructural importance, but do not tend to change over time, thus enabling additional application of this data. Machine learning methods and the application of CNN require extensive work related to preparing the training dataset, which, when it comes to topographic maps, directly depends on the symbology applied in the specific edition of the map. Although the best result of segmenting desired details from the map is expected to be obtained, any changes in the map's symbology mean that the trained model cannot be used for different editions. Therefore, it was decided that in the initial stages of research on data processing and extraction from these maps, it is good to address the problem of detail extraction through Computer Vision methods and color-based segmentation to lay a good foundation for further progress, but also for continued use of these methods when justified.

2. PRESENTATION OF COMMUNICATIONS AT TM25

In order to be able to explain how to detect graphic elements on a map, it is first necessary to explain the symbology of those elements. Geographic elements form the basic part of a topographic map - the map's geographic content [8]. Within them, communications are defined as objects that enable traffic. The elements of land, water, and air transport are shown on the topographic maps. Within land transport, railways and roads are distinguished. Roads can be further classified based on various characteristics. Thus, according to the type of pavement, modern (concrete or asphalt), macadam, and dirt roads meet.

In accordance with the instructions for making TM25 [9] highways, asphalt roads wider than 6 m and asphalt roads from 3 to 6 m can be classified as roads with modern pavement. What all roads with modern pavement have in common is that they are defined in the topographical key by a continuous line symbol in some shade of orange [10], but also that macadam roads are presented with a dashed orange line (Figure 1). As can be seen from the attached image, the symbol for the highway (a) and the asphalt road wider than 6 m (b) have identical shades of color, in contrast to the asphalt road from 3 to 6 m (c) and the macadam road (d). Although it does not belong to the group of modern roads, it is important to define the symbology of the macadam road for elimination in the later stages of processing. These color shades are mostly found only among these symbols, so they are suitable for extraction.

Figure 1 - Symbol for highway (a), asphalt road wider than 6 m (b), asphalt road from 3 to 6 m (c) and macadam road (d).

It is important to note that streets in populated areas are given a special symbol, but that symbol does not carry the information about whether a street is paved or not, but the criterion for its display is the fact that it connects two parts of the settlement. In case a modern road passes through a settlement, the symbol of the street is not drawn, but of the road [11]. In this way, it is possible to follow where the modern road passes through the inhabited place. The digital topographic key DTM25 is based on the topographic key TM25, so the symbology related to the roads has basically remained the same. It can be assumed that the asphalt roads shown on TM25 are also on DTM25, with a certain probability that the road from the category of asphalt road from 3 to 6 has been renewed and widened, and the road is wider than 6 m. It is also safe to assume that modern roads are located in the same position, with the fact that there may be a newly built road or fork in that area.

3. WORK METHODOLOGY

3.1. PREPARING TK25 FOR VECTORIZATION

The preparation of sheets for automatic vectorization is completely done using the *Python* programming language and the *OpenCV* library. The advantage of this approach is that in a very short time, a large number of sheets can be brought into a state suitable for further processing. The procedure begins by loading a georeferenced map sheet. From the map, it is necessary to eliminate all content that does not refer to modern roads. This includes other geographical elements of the map, such as hydrography, objects, vegetation and other elements, but also certain elements of communication, that is, roads whose pavement is made of materials other than asphalt or concrete. In more complex maps, the use of color information is essential for recognizing its features [12]. It has been observed that color is what distinguishes modern roads from other roads. Also, besides the roads, there is very little content on the map that is shown in the same color. These are mostly point symbols that denote objects (mostly religious - synagogues, monasteries, mosques).

The lower and upper limits of the pixel value whose range allows the extraction of orange details are determined empirically. Extracting content based on pixel values proved to be an acceptable solution. However, with this procedure, the macadam roads, which are represented on the maps with a combination of orange and white colors, as well as the mentioned dot symbols (Figure 2, marked with numbers 1 and 2) are retained in the images. The problem itself could be solved by adjusting the Gap Closure Tolerance parameter in *ArcScan* during later processing so that during vectorization the program does not vectorize lines that are at a greater distance than the set one. However, there are not rare situations where, due to the transparency of the map, in some cases, there is a slight overlap of elements. Then it can happen that a part of the symbol, the whole symbol or a textual printout (annotation) can be found on the communication symbol (Figure 2, marked with number 3). Water leaks are given in black color (Figure 2, marked with number 4). All types of bridges and overpasses on modern roads are represented by a combination of black and white colors (Figure 2, marked with number 5). All this leads to the creation of gaps in the array of pixels in the places where they are located. As some gaps are equal to or even larger than the orange sections of the macadam road, it is clear that only adjusting the *Gap Closure Tolerance* parameter will not give satisfactory results. There will be interruptions in the road network in places where there are larger overpasses or bridges, which in reality do not exist.

Figure 2 - View of the map segment (left) and the corresponding extracted imageroads (right) with appropriate markings.

The elimination of macadam roads and point symbols was performed by a combination of the *Laplacian, findContours* and *contourArea* functions from the *OpenCV* library. After detecting the contours of the roads and point symbols that have remained in the image, *contourArea* is used to calculate their areas. The area of contours that should not be in the image was determined experimentally. They are filled with white pixels and issued on a special mask. A mask with unnecessary contours is combined with an image containing only orange symbols, so only white pixels are transferred from the mask to the image. Through the assignment of a new value, the orange sections of the macadam roads and the dotted symbols are converted into white pixels. The procedure was necessary because if only the contours of small areas were erased and the corresponding ones left, the image would not be suitable for obtaining the desired results during automatic vectorization. After this part, the image is freed from unnecessary objects, but there are still gaps in the road network. These gaps were solved using dilation and erosion. Dilation and erosion are the basic morphological transformations, and they arise in a wide variety of contexts such as removing noise, isolating individual elements, and joining disparate elements in an image [13]. The number of iterations and the structuring element parameter were determined experimentally. A structuring element can be simply defined as a configuration of pixels on which an origin is defined, also called an anchor point [14]. In the case of dilatation, it is an ellipse, and in the case of erosion, a square is used. In this way, almost all the gaps were filled. The ones that remained were too far apart. If the number of dilation iterations or the structuring element were to increase, there would be a danger that close parallel paths would merge. A small number of those cases were solved by adjusting the aforementioned *Gap Closure Tolerance* parameter, but this time without the risk of gaps in macadam roads. Raster data prepared in this way are ready for automatic vectorization.

3.2. AUTOMATIC VECTORIZATION USING THE ARCSCAN EXTENSION

Above the previously processed raster map, a map of extracted roads is obtained, on which it is possible to perform automatic vectorization. The automatic vectorization process was performed within *ArcGIS* (version 10.5) software. First, binary raster reclassification was done using the *Reclassify* function. The binary classified image becomes the input data for the *ArcScan* extension within the software, which is intended for the processing and vectorization of raster data. *ArcScan* offers tools to convert scanned images into vector layers. An interactive vectorization experience to draw raster cells on a map to create vector features. The automatic vectorization experience requires feature generation for the entire raster based on predefined settings. As a large amount of geographic information still exists in the form of printed maps, a tool to integrate these documents into GIS is essential [12]. *ArcScan* provides an efficient way to speed up this integration process compared to traditional techniques, such as manual digitization. The *ArcScan* extension also provides tools that allow simple raster editing to prepare layers for vectorization. This practice, known as raster preprocessing, can help remove unwanted raster elements that are not part of the vectorization project. This is exactly what was done in this case in order to further clean up the resulting raster map and reduce the possibility of creating false vector data. On this occasion, all individual elements that are smaller than 50 pixels are deleted. Also, the vectorization style is set to be polygonal. It is important to address certain cases on the map with an adequate selection of parameters. As the topographical content is fitted and overlapped, and some elements are masked, it is necessary to pay particular attention to the *Hole Size* (value 5), *Gap Clouse Tolerance* (value 100) and *Fan Angle* (value 180) parameters in order to ensure that all holes, discontinuities (due to masking, the appearance of bridges and the like) and curves, which were possibly left

behind on the prepared rasters, were treated as a single road. In this way, the vector data of roads that are topologically correct were obtained, so each road segment was generated separately.

The obtained data are raw because they represent the axis of the road, so exported in shapefile format (*.shp*) can be used for other purposes with subsequent definition of symbology as desired. The described process is shown in Figure 3.

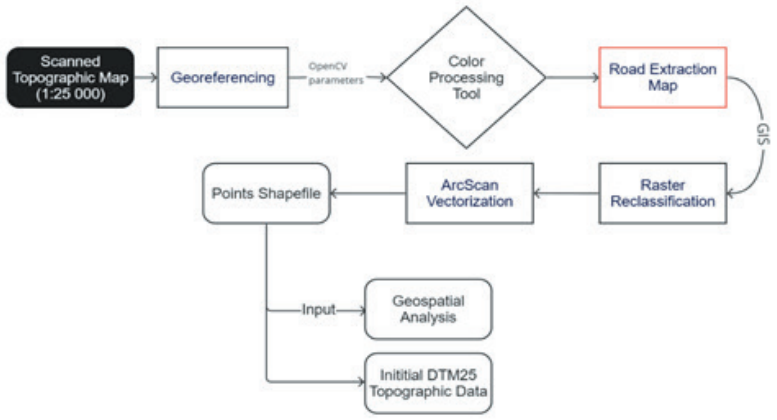

Figure 3 - Workflow chart of applied methodology in work.

4. RESULTS AND DISCUSSION

4.1. DATA USED

The proposed methodology was applied to a sample of 8 TM25. The sheets refer to the territory of the wider region of the city of Kruševac. The selected sheets form a unit of four linked sheets of cards in two rows. The resolution of the scanned sheets is of the order of 4000×5500 pixels and 72 dpi. The state of the contents of the mentioned answer sheets in 1969.

Also, a comparison was made of the obtained data with the existing DTM25 data taken from the MGI *Central Geotopographic Database*, for the territory covered by these sheets, they were field checked and supplemented in the third and fourth quarters of 2022, so it can be concluded that they have a high level of credibility. During the field check of the contents of the topographic map, among other things, the type of pavement and the width of the roads are checked.

4.2. RESULTS OF ROAD VECTORIZATION

In accordance with introductory considerations, the purpose of such results can be multiple. As the territory of the sheets used for input data is already covered by the modern DTM25, a comparison of the state of the road network from 1969 and 2022 shown on the topographic maps will be presented here. Table 1 shows the length of the road network of asphalt roads from 1969 compared to the length of the road network of asphalt roads from 2022 according to TM25 sheets. The expected increase in the length of the road network can be seen, which is particularly noticeable on list 531-4-4 (Makrešane) where the length of the road network has increased by 315.85 %, and the least noticeable on list 531-3-1 (Medveđa) where this increase amounts slightly by 22%.

Table 1 - Showing the comparison of the length of the road network.

List nomenclature	List name	Asphalt roads 1969 (km)	Asphalt roads 2022 (km)	Difference (%)
531-3-1	Medveđa	55.7	68.5	22.98
531-3-2	Velika Drenova-sever	35.9	79.2	120.61
531-3-3	Trstenik	39.2	86.3	120.15
531-3-4	Velika Drenova-jug	51.2	91.8	79.30
531-4-1	Varvarin	32.3	82.3	154.80
531-4-2	Stalać	39.7	85.3	114.86
531-4-3	Kruševac	71.4	146.3	104.90
531-4-4	Makrešane	18.3	76.1	315.85

In general, various analyzes can be performed on the basis of this data. Calculating the density of the road network, comparing economic development with the density of that network or the structure of the roads are just some of the suggestions. A big advantage is that in a very short time, data can be obtained for entire districts, provinces or the entire country. For parts of the territory covered by all editions of topographic maps, they can be analyzed in more detail in terms of changes in the road network over time.

It should be determined whether there is justification for vectorizing other categories of roads from older editions of topographic maps. Macadam and packed stone roads, but above all dirt roads, have a huge tendency to change routes, destroying and creating new ones in short time intervals. This is particularly prominent in Vojvodina, but also in hilly and mountainous areas, mostly in eastern Serbia, where there are fewer and fewer people and the need to maintain dirt roads. Another additional obstacle when extracting all categories of roads using color-based segmentation methods is the fact that there are other linear map elements represented in the same or similar color, as stated in [7] and [15]. As a problem, this can manifest itself even more in the case of the extraction of lower order road categories, which in their composition have colors that are more similar compared to the other content of the map, e.g. dirt roads and grid lines.

4.3. POSITIONAL ACCURACY OF GENERATED LINEAR VECTOR DATA

Regarding the use of the obtained results as a starting point for the creation of a modern road network during the creation of DTM25, it can be concluded that the positional accuracy of the generated data is at a satisfactory level. This is based on the fact that after the georeferenced scanned map is processed, the newly obtained raster has the same image dimensions, so the accompanying files related to the spatial references of the original raster can be used to georeference the processed raster. It is important to note that the conclusion on positional accuracy is made in relation to the existing georeferenced topographic map, that is, the road representation on that map. When it comes to displaying roads on topographic maps, it must be emphasized that these data are cartographically modeled in accordance with the linear scale of symbols and positioned on the map according to certain rules. When the road extends directly next to a water surface or a railway line, due to the scale of the map it is impossible to provide all symbols with the same positional accuracy. In that case, water surfaces, then railroad tracks and finally roads, are given priority in more faithful positioning. It is an insight into the positional accuracy of the data obtained from the topographic map. From this, it can be concluded that the positional accuracy of the obtained roads cannot be higher than the positional accuracy of the roads on the map itself, from which the data is extracted. Another insight is reflected in the matching of the obtained vector data with the content on the georeferenced topographic map, i.e. the accuracy of the vectorized data in relation to the map of the extracted roads. Figure 4 shows map segments, extracted road segments, as well as the final vectorized data compared to the initial map.

Figure 4 - Presentation of methodology steps and positional accuracy of roads (Map segment – left, Road extraction map – Middle, Road vector data – right).

Certain deviations that can be found in certain parts arise as a result of the vectorization of lines that have experienced minor deformations in certain places due to dilation and erosion, as seen in Figure 5. Such cases are seldom and occur with sharp and short curves or when joining roads under small angles. Possible deviations occur at specific road breaks for the purpose of masking and fitting other cartographic content that cannot be addressed by selecting global parameters during vectorization. Also, at the junctions of modern and macadam roads where the beginning of the macadam road at the junction starts with an orange color, it is possible that a false detection occurred on a short segment.

5. CONCLUSION

As the production process of DTK25 is constantly being improved and supplemented with new technological solutions that emphasize an increasing part of the work related to vectorization is transferred from a manual to an automatic work process, with the aim that the obtained geotopographic material is of adequate quality, human resources are freed that are they can engage in other jobs. Also, the development of tools and methods of image processing and computer vision creates the possibility for using the vast amount of data, cartographic materials, and archival materials that have been generated over the long tradition of this institution. The challenge remains to find an approach to process these data so that they are usable for GIS analyses.

The method presented in this paper shows that today it is easier and faster than ever to get this data. It is also focused on several symbols within one geographical element of the topographic map content. When looking at the entire geographic content of a map, it is concluded that there are more thematic units of data that can be processed in similar ways. Topographic data obtained through this automatic process have adequate positional accuracy, and the method is scalable, making it possible to apply it to a large number of topographic maps, thus making the time required to obtain a vector road network corresponding to a past period negligible. The emphasis of the work is on data extraction, rather than on their application and analysis. During the development of the code, as with the vectorization itself, a large number of parameters for the functions used were tested in order to obtain optimal results that can be applied to all sheets.

Further improvements can be made in the domain of positional accuracy of individual parts of vectorized roads by considering the use of an even larger number of parameters and the effectiveness of deep learning methods and training dataset preparation.

6. REFERENCES

[1] M. Petreca and G. Čolović, Geodetska služba JNA, Beograd: Vojnoizdavački i novinski centar, 1987.

[2] R. Banković, S. Tatomirović, D. Đorđević and M. Milašinović, 140 godina vojnogeografskog instituta, Beograd: Medija centar „Odbrana“, 2016.

[3] V. Marković, S. Bakrač, N. Dimitrijević, R. Banković and S. Drobnjak, "Usage of IT in the Process of Topographic Map Creation in MGI," in Sinteza 2020 - International Scientific Conference on Information Technology and Data Related Research, Belgrade, 2020.

[4] M. Basarić, M. Mrlješ and B. Saša , "Point Object Extraction from Scanned Topographic Maps for the Digital Topographic Maps Production," in Sinteza 2022 - International Scientific Conference on Information Technology and Data Related Research, Belgrade, 2022.

[5] J. H. Uhl, S. Leyk, Y. Y. Chiang, W. Duan and C. A. Knoblock, “Automated Extraction of Human Settlement Patterns from Historical Topographic Map Series Using Weakly Supervised Convolutional Neural Networks,” IEEE Access, vol. 8, pp. 6978-6996, 2020.

[6] B. Ekim, E. Sertel and E. Kabadayi, “Automatic Road Extraction from Historical Maps Using Deep Learning Techniques: A Regional Case Study of Turkey in a German World War II Map,” ISPRS International Journal of Geo-Information, vol. 10, no. 8, p. 492, 2021.

[7] C. Jiao, M. Heitzler and L. Hurni, “A survey of road feature extraction methods from raster maps,” Transactions in GIS, vol. 25, no. 6, pp. 2734-2763, 2021.

[8] M. Peterca, N. Radošević, S. Milisavljević and F. Racetin, Kartografija, Beograd: Vojnogeografski institut, 1974.

[9] Vojnogeografski institut, Upustvo za izvođenje radova na II izdanju karte razmera 1:25000, Beograd: Vojnogeografski institut, 1973.

[10] Vojnogeografski institut, Topografski znakovi, Beograd: Vojnogeografski institut, 1981.

[11] Vojnogeografski institut, Upustvo za izradu DTK25, Beograd: Nepublikovano, 2019.

[12] P. Arrighi and P. Soille, “From scanned topographic maps to digital elevation models,” in GeoVision 99: International Symposium on Imaging Applications in Geology, Liège, 1999.

[13] G. Bradski and A. Kaehler, Learning OpenCV, Sebastopol: O'Reilly Media, 2008.

[14] R. Laganiere, OpenCV 3 Computer Vision Application Programming Cookbook, Birmingham: Packt Publishing Ltd., 2017.

[15] R. Samet, I. Askerzade and C. Varol, "An Implementation of Automatic Contour Extraction from Scanned Digital Topographic Maps," Applied and Computational Mathematics, vol. 9, no. 1, pp. 116-127, 2010.

[16] ESRI, "What is ArcScan? - ArcGIS Desktop.," 2021. [Online]. Available: https://desktop.arcgis.com/en/arcmap/latest/extensions/arcscan/what-is-arcscan-.htm#:~:text=ArcScan%20provide-.htm#:~:text=ArcScan%20provides%20tools%20that%20allow,automatically%20using%20the%20automatic%20mode. [Accessed 4 May 2023].