International Journal of Engineering and Geosciences

https://dergipark.org.tr/en/pub/ijeg

e-ISSN 2548-0960

A new approach for matching road lines using efficiency rates of similarity measures

Müslüm Hacar*[1], Turkay Gökgöz[1]

[1]*Yıldız Technical University, Engineering Faculty, Geomatics Engineering Department, Istanbul, Turkey*

Keywords
Object matching
Similarity measure
Efficiency ratio
Geometric integration
Map conflation

ABSTRACT

The lack of common semantic information among corresponding geo-objects in different datasets required new matching approaches based on geometric and topological measures. In this study, a semi-automated matching approach based on the matching capabilities of geometric and topological measures was proposed. In the first stage, after the initial matching performed by a scoring system, the efficiency of each measure on the matching accuracy is evaluated manually by an operator. In the second stage, (1) the score of each measure is updated in accordance with the accuracy distributions. This means that the score of a measure is increased if it is relatively more significant than others. Finally, (2) matching process is repeated with new scores. The proposed approach was tested by matching tree-, cellular-, and hybrid-patterned road lines in municipal, private navigation, and OpenStreetMap datasets. The experimental testing shows that it has satisfactory results both in accuracy and completeness. F-measure is over 86% in hybrid-patterned Bosphorus datasets.

1. INTRODUCTION

Geometric integration establishes the relationships between the objects in a spatial dataset and the corresponding objects in another dataset and ensures that the target dataset reaches the required competence. Producing better (geometrically and semantically more up-to-date and rich) maps by using two different maps representing the same entities is also an important issue of integration and is called map conflation (Lynch and Saalfeld, 1985; Saalfeld, 1988). The integration process can be used for different purposes. Cobb et al. (1998) remarked the requirements for map conflation as; updating with the objects transferred from one dataset to another, optimization of geometric and semantic accuracy, and transferring data to a dataset containing missing information. The conflation process enables the spatial data generated by different sources to be used together. Geometric, topological and semantic similarities between objects are important criteria for the conflation process. The greater the similarity, the lower the operator effort.

The conflation process is based on the principle of matching geometries (point, line, and polygon) that represent the same real entities (Yuan and Tao, 1999). Determining the correspondences between the objects according to their relations and similarities is called matching process (Hacar and Gökgöz, 2019b).

In this study, a semi-automated matching approach based on the efficiency rates of the measures was proposed. In this section, related studies in the literature are examined. Following section presents the study area and datasets, the geometric and topological measures used to determine the similarities between the objects, and the proposed approach. Section 3 presents the experiment with tree-, cellular-, and hybrid-patterned road networks, and the evaluation of the results conducted with the statistics of the study. Section 4 concludes the study by discussing the results and giving several further suggestions.

1.1. Related Works

Many methods have been developed to match line objects since it was first applied in 1980s by Rosen and Saalfeld (1985) and Saalfeld (1988). Main problem in line matching is that none of the corresponding line geometries from different sources are geometrically identical. In other words, the geometrical properties of corresponding line objects such as orientation, length, shape, location have not equal values. According to Hacar and Gökgöz (2019b), there are three important reasons that researchers prefer to work with line matching rather than point and polygon matchings: (1) difficulties in establishing relationships between complex representations such as patterns, intersections,

* **Corresponding Author**

*(muslumhacar@gmail.com)ORCID ID 0000-0002-8737-8262
(gokgoz@yildiz.edu.tr) ORCID ID 0000-0001-8716-6131

Cite this article

Hacar M & Gökgöz T (2021). A new approach for matching road lines using efficiency rates of similarity measures. International Journal of Engineering and Geosciences, 6(3), 146-156

Research Article / DOI: 10.26833/ijeg.791324

Received: 07/09/2020; Accepted: 04/11/2020

roundabouts, dead ends, (2) the need to keep navigation datasets up-to-date and (3) the rise of Volunteered Geographical Information (VGI) datasets.

The concepts of matching progress in spatial data integration have also been focused by researchers. Yuan and Tao (1999) classified the matching process by geometry, topology and semantic. Ruiz et al. (2011) also discussed the integration process by match type; geometric, topological and semantic. Volz (2006) classified the process by similarity measures; point-, linear-, and area-based and the hybrid. Xavier et al. (2016) classified the measures as geometric, topological, attribute, context, and semantic. Memduhoğlu and Başaraner (2018) compared thematic geographic ontologies created for cities and discussed about possible contributions of basic integration methods and technologies of spatial semantics for creating a multi-representation spatial database paradigm. Hacar and Gökgöz (2019a) designed a conceptual model for matching process under spatial data integration by classifying the types of geometry, measure, relationship, and spatial information.

There have been developed many matching methods. While some of them works fully automated, others allow the user intervention. Xiong and Sperling (2004) proposed a semi-automatic method for matching road networks. By using a cluster-based matching process, strong relationships between nodes, edges, and segments in the two road networks are determined. Their method allows identifying and correcting missing matches, but requires significant interaction (operator intervention) during the process. Li and Goodchild (2011) proposed an automated optimization model to match the road lines using geometric and semantic measures, as well as an affine transformation. They used asymmetric property of one-way Hausdorff distance as a measure of dissimilarity. In addition, the Hamming distance was also used as a criterion of dissimilarity to show the difference between road names. Lei and Lei (2019) also developed a flow-based optimization model that seeks to minimize the total discrepancy between two datasets. Moreover, Araújo et al. (2019) proposed a Spark-based approach using the names of the places (semantic) and context information (e.g., neighbouring streets) to compare the corresponding objects in real-world data sources of New York and Curitiba.

Some researchers focused on matching objects in datasets that have a significant scale difference. To work with this kind of source datasets, researchers often use topological measures (e.g., the degree of connectivity (or the valence), spider function, buffer-growing, etc.) to match the corresponding objects. Mustière and Devogele (2008) proposed an approach relying on the comparison of geometrical, attributive, and topological properties of objects for matching networks with different levels of details. Olteanu-Raimond et al. (2015) used belief theory to represent and fuse knowledge from different sources to model imperfection (imprecision, uncertainty, and incompleteness), and make a decision. Chehreghan and Abbaspour (2018) developed an optimization-based matching approach for multi-source spatial datasets by taking into account several geometric criteria. The approach benefits from a genetic algorithm and sensitivity analysis to identify corresponding objects. Moreover, Guo et al. (2019) designed a new matching method for the objects in multi-scale geodatabases using weights of some well-known geometric and topological measures. The method has three stages; (1) entire, (2) partial matchings, and (3) roundabout detection and matching. The authors used a splitting process to match the unmatched road segments.

Some studies in urban lands are also crucial tasks of integration cases. Recently, VGI, social media, and geocoding data are used to extract and combine new spatial data in urban areas (Hacar 2020; Kılıç and Gülgen 2020; Bilgi et al. 2019). VGI enables generation of maps by using crowd-sourced volunteer contributors. Each volunteer has equal role to contribute the geometric and semantic properties of the geographical objects. However, since there is no rule to be a volunteer in VGI, non-expert contributors may draw features irregularly or inconsistently with basics of cartography. Therefore, result map may have low quality. In this context, geo-object matching is used as a process providing a solution for analysing and increasing the quality and accuracy of VGI data. Koukoletsos et al. (2012) proposed a matching approach to assess the completeness of VGI data. They developed a multi-step approach matching OpenStreetMap (OSM) road data with the UK's official mapping agency Ordnance Survey (OS), taking into account the similarities in geometric (search distance, direction, line-based buffer zone) and attribute (road names). Pourabdollah et al. (2013) also conducted a conflation study with attribute-rich OS data to improve the quality of OSM road data. Besides, Hacar and Gökgöz (2019b) conducted a matching study with OSM and TomTom navigation data. In some cases, line-based (linear) approaches to matching road objects may be insufficient. In such cases, an area-based (spatial) matching approach, like proposed by Fan et al. (2016), can be used. This method finds the corresponding blocks in source datasets with a spatial overlapping ratio. It then matches the surrounding roads using the matched blocks. Also Fan et al. (2016) tested their method by matching OSM and public city data and achieved satisfactory results in Heidelberg (Germany), a network of regular networks, and Shanghai (China), with a relatively more complex network. The sources and patterns of road networks are two important factors to consider in the matching process. Yang et al. (2014) classify the pattern groups of the blocks that the roads surround and match the nodes in the groups hierarchically. Hacar (2019) and Hacar and Gökgöz (2019b) developed a score-based multi-stage method and tested it with cellular-, tree-, and hybrid- patterned road networks. According to the method, the candidate matches are scored in accordance with the geometric and topological similarity and then the objects with high scores are matched incrementally.

The matching methods differ from each other according to the hierarchical steps of the approaches, even if they have some common stages, metrics or rules. The design of the method can primarily affect the sufficiency of the case study. Also, the complexity of road networks can reduce the sufficiency. The previous approaches had low interest in complex road networks

such as in Istanbul. In this study, the scope of the proposed approach is determined to design a new matching model and its applicability in Istanbul road networks.

2. THE PROPOSED APPROACH

The proposed approach performs the matching process of road lines thanks to the efficiency rates. The rates are calculated using geometric and topological measures. The main idea for selecting the measures is to determine the similarities of corresponding matching pairs from different source datasets. As seen in Fig. 1, the matching process is managed in two stages in addition to a pre-process. Firstly, two road networks are aligned as a pre-process. In the first stage, road lines closer to each other than a predefined threshold distance value T are identified as candidate matchings. Hausdorff distance is used to determine the closeness between candidates. T should be large enough to identify possible correct matches and small enough not to cause too many missing matches (mismatching). The threshold can be determined by examining the source datasets and structure of road networks, and by conducting several experimental matching observations. After the selection of corresponding pairs, for each candidate matching, (1) similarity scores are calculated based on the measures of Hausdorff distance (S_H), orientation (S_O), sinuosity (S_S), mean perpendicular distances (S_P), mean length of triangular edges (S_T) and modified degree of connectivity (S_C) (Fig. 2). The maximum similarity score assigned to a candidate pair is 4 for all measures apart from sinuosity and mean perpendicular distance. Sinuosity and mean perpendicular distance represent similar characteristics of lines. The maximum score with respect to these indicators is 2 so that the maximum total score of these indicators shall be the same as the others, 4, for fairness (Hacar and Gökgöz, 2019b). Table 1 shows the computation criteria of scores for each measure. (2) Sum of the similarity scores are obtained for each candidate pair, and (3) the candidates, whose total similarity scores are maximum, are selected as matched pairs and other candidates are eliminated. The efficiency of each measure is determined by comparing the matched pairs with the result of the manual matching. After determining the number of correct and incorrect matches for each measure, it is ensured that the score of the measure, which performs better results in term of the number of correct and incorrect matches, is higher than that of the relatively insignificant (less number of correct and/or much more incorrect matches). For this purpose, the efficiency ratio is used, where the numbers of correct and incorrect matchings are placed together. Each measure has its own efficiency ratio.

Maximum-Minimum normalization method was adapted to calculate the efficiency ratio. Briefly, the ratio is multiplied by the similarity scores to increase the effect of the measure that performs the matching process with high accuracy and reduces, but not disables, the effect of the measure with low accuracy.

The normalization consists of two equation: *Profit* (P) and *Loss* (L) (Eq. 1 and Eq. 2). While P represents how far the value X_i from minimum value, L represents how close the value X_i to maximum value. The following formulas are used as original Maximum-Minimum normalization measures (Başaraner, 2011; Şen, 2013).

$$P = \frac{X_i - X_{min}}{X_{max} - X_{min}} \tag{1}$$

$$L = \frac{X_{max} - X_i}{X_{max} - X_{min}} \tag{2}$$

These criteria can be adapted to calculate the normalized values P_i and L_i for each similarity measure with regards to the correct and incorrect match numbers as follows.

$$P_i = \frac{N_{Correct_i} - N_{Correct_{min}}}{N_{Correct_{max}} - N_{Correct_{min}}} \tag{3}$$

$$L_i = \frac{N_{Incorrect_{max}} - N_{Incorrect_i}}{N_{Incorrect_{max}} - N_{Incorrect_{min}}} \tag{4}$$

where $N_{Correct_i}$ (i=1,2,..,n) represents the number of correct matches of the respective measure, $N_{Correct_{min}}$ represents the least number of correct matches, and $N_{Correct_{max}}$ represents the maximum number of correct matches between all the measures. In addition, $N_{Incorrect_i}$ (i=1,2,..,n) represents the number of incorrect matches of the respective measure, $N_{Incorrect_{min}}$ represents the least number of incorrect matches, and $N_{Incorrect_{max}}$ represents the maximum number of incorrect matches between all the measures.

The efficiency rates could be calculated as follows:

$$E_i = P_i \times L_i \tag{5}$$

However, the efficiency ratio (Eq. 5) is to be zero for the measure that performs the maximum number of incorrect or minimum number of correct matches. This results in the score used for the respective measure being multiplied by a factor of 0 (zero) and the corresponding measure being ineffective (disabled) in the second stage of the approach. Since there is no correlation between the numbers of the correct and incorrect matches, making any measure ineffective may reduce the success of the process. Also, our experience in matching cases motivates us to consider all of the measures, even if it is relatively less significant (generating many incorrect matches). Therefore, the exponential function should be used with previous formula (Eq. 5). Exponential function prevents the least important measure from taking a value 0 (Eq. 6). In other words, the least important measure also affects the results in the second stage.

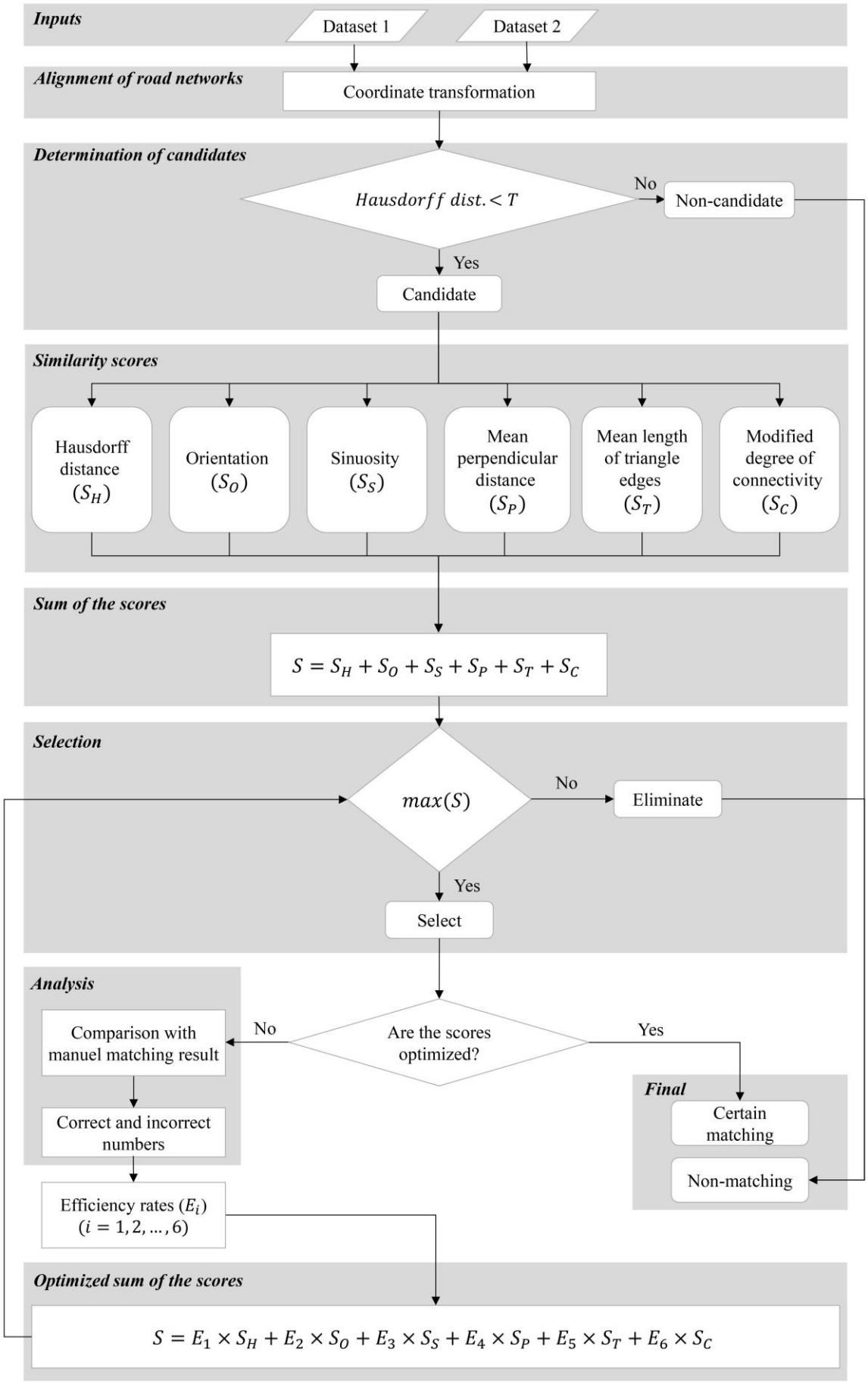

Figure 1. The workflow of the proposed approach

Table 1. The computation criteria of similarity scores (Hacar, (2019); Hacar and Gökgöz, 2019b)

Measure	Criteria
Hausdorff distance	For each candidate pair, the first three closest matches are scored as $S_{H_1} = 4$, $S_{H_2} = 2$, and $S_{H_3} = 1$, respectively. The fourth and others are scored as $S_{H_{i\,(i>3\,\in Z^+)}} = 0$.
Orientation	Candidate pairs in the same class are scored as $S_O = 4$. If they are in adjacent classes (seen in Fig. 2)), the score is assigned as $S_O = 2$. Otherwise, the score is assigned as $S_O = 0$.
Sinuosity	The rules for sinuosity scores (S_S) are as follows: if S_n = Low and if S_m = Low, then $S_S = 2$ if S_n = Low and if S_m = Mid, then $S_S = 1$ if S_n = Low and if S_m = High, then $S_S = 0$ if S_n = Mid and if S_m = Low, then $S_S = 1$ if S_n = Mid and if S_m = Mid, then $S_S = 2$ if S_n = Mid and if S_m = High, then $S_S = 1$ if S_n = High and if S_m = Low, then $S_S = 0$ if S_n = High and if S_m = Mid, then $S_S = 1$ if S_n = High and if S_m = High, then $S_S = 2$
Mean perpendicular distance	If the difference between the mean perpendicular distances of Line n and Line m is less than or equal to $\sigma_p/2$ (σ_p is the standard deviation of all mean perpendicular distances), then it is scored as $S_P = 2$. If the difference between the mean perpendicular distances of Line n and Line m is greater than $\sigma_p/2$ and less than or equal to σ_p, then it is scored as $S_P = 1$. Otherwise, it is scored as $S_P = 0$.
Mean length of triangle edges	If the difference between the mean length of triangle edges of Line n and Line m is less than or equal to $\sigma_E/2$ (σ_E is the standard deviation of all mean lengths of triangle edges), then this matching is scored as $S_T = 4$. If the difference between the mean length of triangle edges of Line n and Line m is greater than $\sigma_E/2$ and less than or equal to σ_E, then it is scored as $S_T = 2$. Otherwise, it is scored as $S_T = 0$.
Modified degree of connectivity	If the candidates have the same degree, then it is scored as $S_C = 4$. If there is a just one degree of difference between the candidates, then it is scored as $S_C = 2$. Otherwise, it is scored as $S_C = 0$.

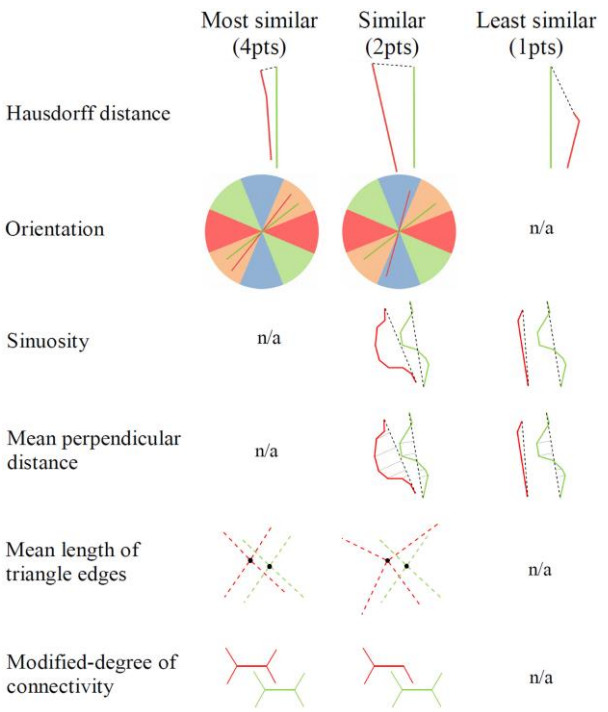

Figure 2. The similarity scores of possible matches (Hacar and Gökgöz, 2019b).

$$E_i = 2^{P_i \times L_i} \tag{6}$$

In the second stage, the matching process is repeated with similarity scores updated (optimized) with E_i efficiency rates. This means that the score of a measure is increased whether it is relatively more significant than others. Finally, the candidates with the highest total similarity scores are determined as certain matches.

3. EXPERIMENTAL TESTING

3.1. Study Area and Datasets

The proposed method was tested with tree-, cellular-, and hybrid-patterned road networks in Istanbul. We used different sources as; Istanbul Metropolitan Municipality (IMM), two private navigation companies Başarsoft and TomTom, and OSM, one of the popular VGI projects, to show how efficient the proposed approach with different samples (Fig. 3) (Table 2). Also, an additional matching process was conducted with a large amount of data covering Bosphorus of Istanbul to prove its efficiency in a realistic way (Fig. 4). In Bosphorus, major elevation differences exist from coastal land to exterior bound. This kind of local surface changes makes road networks complex and leads the road shapes to be similar with hybrid-patterns.

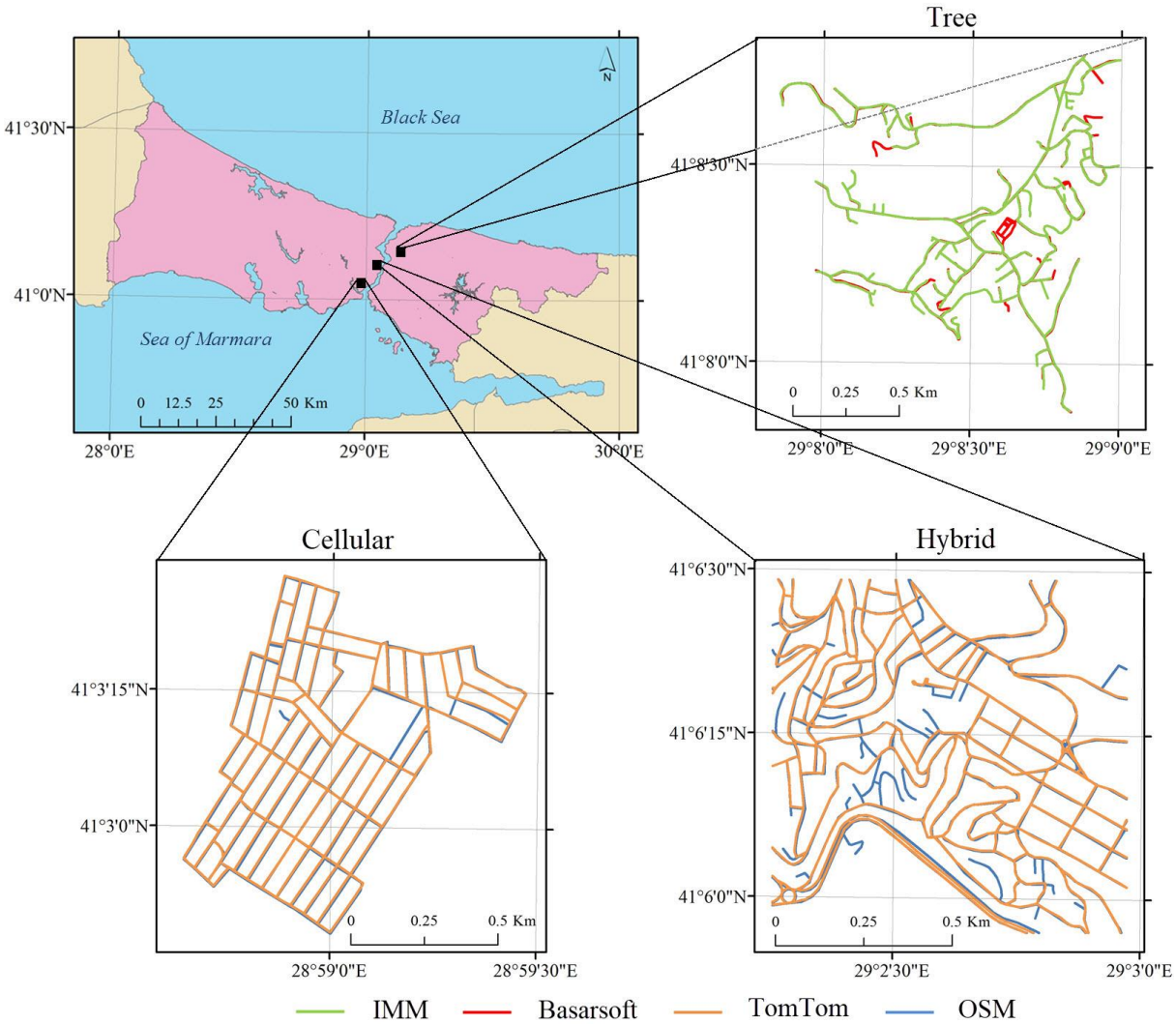

Figure 3. Tree-, cellular-, and hybrid-patterned road networks: IMM (green), Başarsoft (red), OSM (blue), and TomTom (orange)

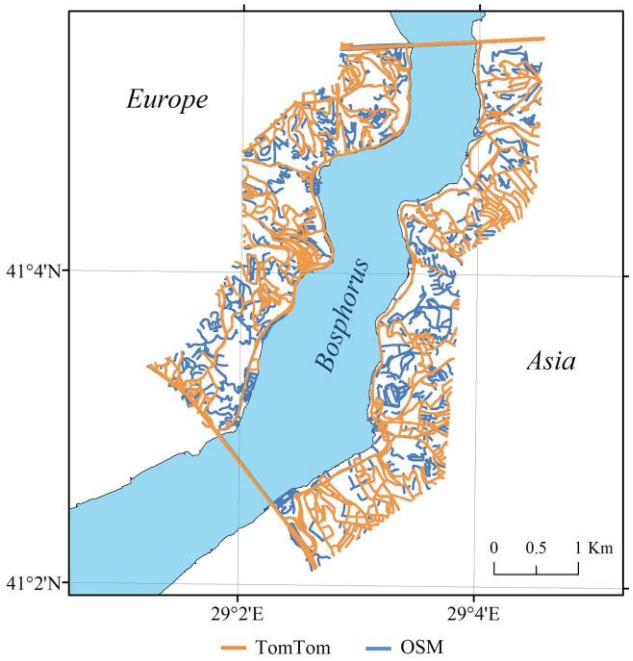

Figure 4. The road networks in Bosphorus, Istanbul: OSM (blue) and TomTom (orange)

Table 2. The number or objects and total road length in each dataset

Pattern	Source	The number of objects	Total length (km)
Tree	IMM	134	14.54
	Başarsoft	118	13.64
Cellular	OSM	153	16.22
	TomTom	146	15.87
Hybrid	OSM	288	21.95
	TomTom	221	19.69
Bosphorus	OSM	3030	221.60
	TomTom	1381	141.04

3.2. Pre-processing

The source datasets have different coordinate systems. This difference affects the calculation of similarity negatively. For example, the objects in the Başarsoft, TomTom, and OSM datasets have geographical coordinates in WGS84 datum. However, the measures used in the study are calculated in metric. Therefore, the geographical coordinates of the objects were transformed into the ITRF96 datum (Gauss-Krüger projection, Central meridian: 30° and GRS80 ellipsoid)

where the IMM dataset was defined. Furthermore, two road networks were aligned using linear rubber-sheet transformation. Moreover, we set T distance threshold as 85m for tree- and cellular-patterned road networks and 50m for hybrid-patterned road network by using our previous matching experiences with source datasets and the study area.

3.3. Results and Evaluation

The results of the matching process were compared with the results of manual matching, and then, the numbers of correct and incorrect matches in Table 3 were determined. The evaluation was performed both integrated and separately with each geometric and topological measure. In the first stage of the approach, some results occurs categorically in accordance with the type of measures and road patterns. While Hausdorff distance measure performed the maximum number of correct and the least number of incorrect matches in both tree and cellular patterns, its result in hybrid pattern is different. Mean perpendicular distance performs the maximum number of correct matches. However, it also gave the most number of incorrect matches in hybrid patterns. Therefore, we examine the results of the measures by using their correctness and incorrectness percentages (Table 3). Hausdorff distance measure performed the maximum correctness and the minimum incorrectness in all patterns. Sinuosity and mean perpendicular distance measure gave the least correctness and the maximum incorrectness in cellular pattern. Orientation was the second best similarity measure in terms of both correct and incorrect matching in all patterns. From this point of view, it can be observed from Table 3 that mean perpendicular distance was the worst in all patterns. Similarly, mean length of triangle edges and modified degree of connectivity performed the least correctness and the most incorrectness in hybrid pattern. However, these measures gave similar results with orientation and sinuosity in tree.

The similarity scores used in the first stage were optimized by the E_i in Table 4 and new similarity scores to be used in the second stage were calculated as in Table 5.

Table 3. The numbers and percentages of correct and incorrect matching

			H[1]	O[2]	S[3]	P[4]	T[5]	C[6]	1. stage
Tree	Correct	Number	90	88	84	83	86	86	88
		%	78	54	49	40	43	47	75
	Incorrect	Number	26	75	86	125	114	97	30
		%	22	46	51	60	57	53	25
Cellular	Correct	Number	146	146	109	144	142	145	147
		%	95	39	33	33	37	39	94
	Incorrect	Number	7	233	224	299	246	231	9
		%	5	61	67	67	63	61	6
Hybrid	Correct	Number	191	191	182	195	181	189	190
		%	83	56	47	40	40	40	82
	Incorrect	Number	38	148	206	296	275	279	42
		%	17	44	53	60	60	60	18

[1]Hausdorff distance; [2]Orientation; [3]Sinuosity; [4]Mean perpendicular distance; [5]Mean length of triangle edges; [6]Modified degree of connectivity

Table 4. Efficiency rates (E_i) of similarity measures

E_i	H[1]	O[2]	S[3]	P[4]	T[5]	C[6]
Tree	2.00000	1.28409	1.03978	1.00000	1.03356	1.08765
Cellular	2.00000	1.16961	1.00000	1.00000	1.11875	1.17006
Hybrid	1.64067	1.32845	1.01742	1.00000	1.00000	1.02644

[1]Hausdorff distance; [2]Orientation; [3]Sinuosity; [4]Mean perpendicular distance; [5]Mean length of triangle edges; [6]Modified degree of connectivity

Table 5. The similarity scores used in the first and second stages of the approach

Stage	Pattern	S_H			S_O			S_S			S_P			S_T		S_C	
1.	All	4	2	1	4	2	2	1	2	1	4	2	4	2			
2.	Tree	8	4	2	5.14	2.57	2.08	1.04	2	1	4.13	2.07	4.35	2.18			
	Cellular	8	4	2	4.68	2.34	2	1	2	1	4.48	2.24	4.68	2.34			
	Hybrid	6.56	3.28	1.64	5.31	2.66	2.03	1.02	2	1	4	2	4.11	2.05			

In the second stage, the relationships between the candidates were determined with new similarity scores in Table 5 and the process was performed for the last time. Accordingly, while the proposed approach, with

the updated (optimized) scores, performed almost the same number of matches as the number of manual matching in tree and cellular patterns, some missing matching occurred in hybrid pattern (Table 6). The missing matching is related to two parameter: (1) matching capability of the approach and (2) distance threshold. While the approach was common for all the source patterns, the distance threshold *T* was different in hybrid pattern. Therefore, possible reason for the missing matches of hybrid was *T*.

With the updated similarity scores, the number of correct matches increased by 4 and the number of incorrect matches decreased by 7 in tree-patterned roads. Although the number of incorrect matches decreased by 2 in cellular-patterned roads, the number of correct matches also decreased by 1. While there is no change in the number of correct matches in hybrid roads after second stage, the number of incorrect matches decreased by 8.

The operation of controlling the manual matching could have been too hard with over a thousand corresponding matching pairs in Bosphorus datasets. Therefore, after generating the final matching with whole datasets, the correct and incorrect matches was determined by comparing randomly selected sample data with manual matching (Fig. 5). In Table 6 and 7, the results are based on the sample of Bosphorus datasets.

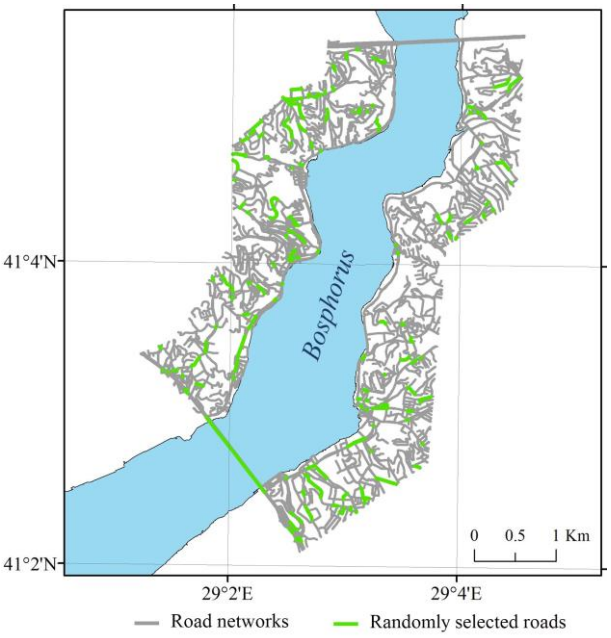

Figure 5. Randomly selected roads (green) and the whole road networks (grey) in Bosphorus datasets.

Since Bosphorus datasets consist of several types of patterns, it is better to examine matching instances in accordance with the pattern type separately. Fig. 6 shows correctly matched road lines with cellular pattern. They were matched correctly both in the first stage and the second stage. Both this visual instance and Table 6 show that the second stage of the proposed method almost have the same result with the first one in cellular patterned-road networks. Besides, while the northwest roads with hybrid pattern in Fig. 7 was matched correctly, the south was a missing match. The possible reason is that the corresponding roads have quite different geometric properties such as sinuosity and centroid. Moreover, the road 1 in Fig. 8 was matched incorrectly with the road 2' both in the first and second stage since the geometric and topological properties of the road 1 are more similar with the road 2' than with the road 1'. As a matter of course, there were expected instances showing us that the second stage optimized the matching process by eliminating the incorrect matches in the first stage. The road 1 Fig. 9b was matched with three roads in other datasets in the first stage. However, the matches with the roads 2' and 3' were incorrect. In the second stage, the efficiency rates ensured the elimination of the incorrect matches.

Table 6. Final results of the matching process by means of matching numbers

		Cor.[1]	Incor.[2]	Miss.[3]	Sum
Tree	Man.[4]	116	-	-	116
	1.Stage	88	30	-	118
	2.Stage	92	23	1	115
Cellular	Man.[4]	150	-	-	150
	1.Stage	147	9	-	156
	2.Stage	146	7	-	153
Hybrid	Man.[4]	262	-	-	262
	1.Stage	190	42	30	232
	2.Stage	190	34	38	224
Bosphorus sample (Hybrid)	Man.[4]	151	-	-	151
	1.Stage	114	25	12	139
	2.Stage	114	18	19	132

[1]Correct; [2]incorrect; [3]missing; [4]manual matching

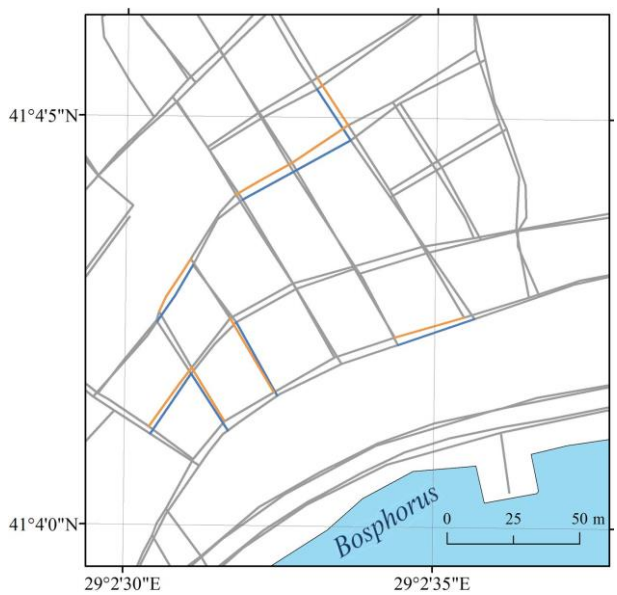

Figure 6. Correct matches in the cellular pattern

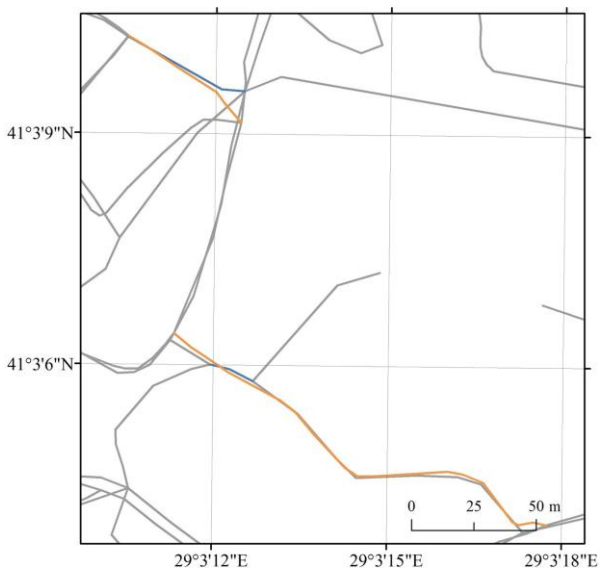

Figure 7. Correct (northwest) and missing (south) matches

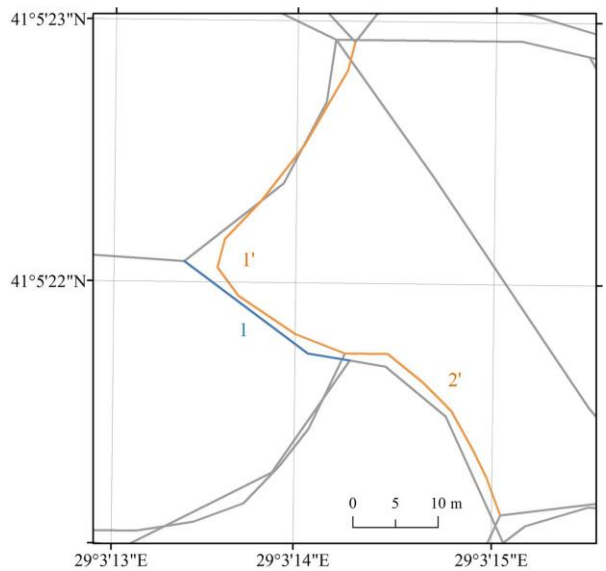

Figure 8. Incorrect matches in the hybrid pattern

Determining the accuracy of a matching study only by the correct matches is not sufficient. For example, in a study area, there are 100 manually detected possible matches and a selected automated method performed 10 matches only. If none of the 10 matches is incorrect, the method is considered to have worked with 100% correctness. However, according to manual matching, the method could not identify 90 matches. This shows that completeness should also be taken into account when making assessments of accuracy. Therefore, three of the frequently used measures of statistical analysis in data science; *precision* (Eq. 7), *recall* (Eq. 8) and *F-measure* (Eq. 9) were used to evaluate the proposed method (Samal et al., 2004; Song et al., 2011; Fan et al., 2016).

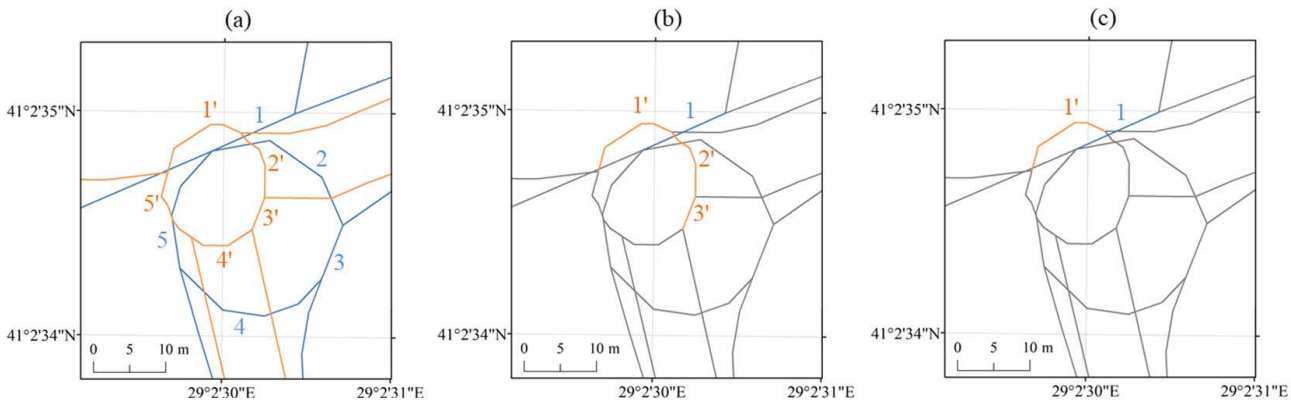

Figure 9. Manual (a), the first stage (b), and the second stage (c) matching

$$Precision = \frac{N_{Correct}}{N_{Correct} + N_{Incorrect}} \quad (7)$$

$$Recall = \frac{N_{Correct}}{N_{Correct} + N_{Mismatch}} \quad (8)$$

$$F - measure = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (9)$$

Three parameters have been used in the statistical measures: Number of correct matches (true positive), number of incorrect matches (false positive), and number of missing matches (false negative) ($N_{Mismatch}$). The number of missing matches was obtained by subtracting the total number of matches performed by the method (sum of true and false matches) from the number of manual matches.

The *precision* measure is a ratio of the number of correct matches to the total number of matches. Therefore, the *precision* was used as the accuracy indicator. *F-measure* is an evaluation measure in which the *precision* (accuracy) and *recall* (completeness) together affect in a balanced way. In the second stage of the method, the accuracy increased by 5.4%, 1.2%, 2.9%, and 4.4% in tree-, cellular-, and hybrid-patterned roads and Bosphorus sample, respectively (Table 7). It can be said that the results are satisfactory in terms of accuracy. *Recall* is a measure of how complete the methods are performed. For instance, when Table 6 is examined carefully, comparing with the manual matching result, the proposed approach performed two more matchings (over-matches) in the first stage and one missing in the second stage with tree-patterned roads. As seen in Table

7, the completeness is 100% in the first stage and 98.9% in the second stage. This means that over-matches do not affect the value of the *recall* measure. This also indicates that the *recall* value cannot be a standalone measure for the evaluation, but can be used to interpret the accuracy. From this point of view, recall value presented that the proposed approach ensured high completeness (almost fully complete). Therefore, the accuracy of the study is quite reliable. In hybrid-patterned roads, the recall value decreased in the second stage. This is because of that while the number of incorrect matches decreased, the number of missing matches increased. Also, F-measure increased by 3.1% and 0.7% in tree and cellular patterns. It has no change in hybrid pattern since (1) the number of correct matches had no change, and (2) decreasing number of incorrect matches was added to number of missing matches in both stages.

Table 7. The results of the evaluation measures

		Prec.[1] (%)	Rec.[2] (%)	F-m.[3] (%)
Tree	1.Stage	74.6	100	85.4
	2.Stage	80.0	98.9	88.5
Cellular	1.Stage	94.2	100	97.0
	2.Stage	95.4	100	97.7
Hybrid	1.Stage	81.9	86.4	84.1
	2.Stage	84.8	83.3	84.1
Bosphorus sample (Hybrid)	1.Stage	82.0	90.5	86.0
	2.Stage	86.4	85.7	86.0

[1]Precision; [2]recall; [3]f-measure

The number of correct matching of each measure is close to each other (Table 3). Therefore, the correct matching numbers have no specifics. This assessment supports the proposed efficiency formula in which the incorrect matches are used. Moreover, Hausdorff distance performed the number of correct matches at least 3.5 times greater than the number of incorrect matches (Table 3). Other measures performed many incorrect matches. Sinuosity and mean perpendicular distance performed the worst in cellular pattern since most of the corresponding road lines has low curvature. The results show that some of the similarity measures are more important than others for the pattern type on which they are used. For instance in our experiments, while Hausdorff distance was the best-matcher for all patterns, the mean length of the triangle edges was the worst-matcher for only hybrid pattern. This kind of changeable order between measures clearly supports the proposed approach that optimizes the similarity scores using the efficiency rates.

4. CONCLUSION

This paper proposes a semi-automated approach for road objects in line geometry. Besides, since it determined the efficiency rates for the tree-, cellular-, and hybrid-patterned road network datasets, the second stage of the proposed approach can be performed automatically with the road networks in a similar pattern. For a road network with a different pattern, the efficiency rates must be recalculated since the similarity measures have different correctness and incorrectness in terms of the pattern type (Table 3). In addition, efficiency rates can be calculated using small samples for datasets containing a large number of road objects, and then, applied to the source datasets. In this case, after the efficiency rates are determined semi-automatically by a manual matching operator using randomly selected samples, the actual large data is matched automatically using these efficiency rates. To prove the efficiency of the proposed approach, we conducted an additional matching process with OSM and TomTom road networks in Bosphorus, Istanbul. Since the Bosphorus networks were hybrid-patterned, the efficiency rates had no need to be computed again. This enables the matching process with the same patterned roads to start directly from the second stage.

Utilization of Maximum-Minimum normalization and the exponential function enabled the efficiency rates to be ranged between 1 and 2. Thus, even the mean perpendicular distance was used as the least significant measure in the similarity calculation.

The proposed approach does not use any semantic information to determine the similarity between objects. Instead, the similarities are calculated on the basis of scores based on geometric and topological measures. The optimization process updates the scores using the efficiency rates.

In this study, the scoring rules and the geometric and topological measures were taken from the study of Hacar and Gökgöz (2019b). However, the proposed approach can be used to adapt different kind of scoring rules using different geometric and topological measures that are specific to the characteristics of the source datasets.

The proposed approach has an *F-measure* over 86% in hybrid-patterned Bosphorus datasets. The results are satisfactory in terms of accuracy and completeness. The experimental testing also show that there is no need to conduct a second stage for the cellular-patterned road networks.

Computing the time of the matching process is a hard task since the process is conducted semi-automatically. The process time changes according to the experiences of the matching operator in the stage of manual results. This may occur the disadvantage that prevents planning the geo-process routines.

ACKNOWLEDMENT

The authors would like to thank IMM Directorate of Geographical Information Systems, The Traffic Stats Customer Service Team in TomTom, and Basarsoft Information Technologies Inc. for supplying road datasets and OpenStreetMap community for their contributions.

REFERENCES

Araújo T, Pires C, Mestre D, de Queiroz A, Santos V & da Nóbrega T (2019). A Parallel-based Map Matching Approach over Urban Place Records. Anais do XXXIV Simpósio Brasileiro de Banco de Dados, 121-132. Porto Alegre: SBC.

Başaraner M (2011). A zone-based iterative building displacement method through the collective use of Voronoi tessellation, spatial analysis and

multicriteria decision making. Boletim de Ciências Geodésicas, 17(2), 161-187.

Bilgi S, Gulnerman A G, Arslanoğlu B, Karaman H & Öztürk Ö (2019). Complexity measures of sports facilities allocation in urban area by metric entropy and public demand compatibility. International Journal of Engineering and Geosciences, 4(3), 141-148.

Chehreghan A & Ali Abbaspour R (2018). A geometric-based approach for road matching on multi-scale datasets using a genetic algorithm. Cartography and Geographic Information Science, 45(3), 255-269.

Cobb M A, Chung M J, Foley III H, Petry F E, Shaw K B & Miller H V (1998). A rule-based approach for the conflation of attributed vector data. GeoInformatica, 2(1), 7-35.

Fan H, Yang B, Zipf A & Rousell A (2016). A polygon-based approach for matching OpenStreetMap road networks with regional transit authority data. International Journal of Geographical Information Science, 30(4), 748-764.

Guo Q, Xu X, Wang Y & Liu J (2019). Combined Matching Approach of Road Networks under Different Scales Considering Constraints of Cartographic Generalization. IEEE Access, 8, 944-956.

Hacar M (2019). Yol Ağlarının Geometrik Entegrasyonu için Nesne Eşleme Yöntemlerinin Geliştirilmesi. PhD Thesis, Yildiz Technical University, Istanbul.

Hacar M (2020). A rule-based approach for generating urban footprint maps: from road network to urban footprint. International Journal of Engineering and Geosciences, 5(2), 100-108.

Hacar M & Gökgöz T (2019a). A conceptual model for geo-object matching. International Symposium on Applied Geoinformatics (ISAG-2019), 7-9 November 2019, Istanbul, Turkey, 97-102.

Hacar M & Gökgöz T (2019b). A New, Score-Based Multi-Stage Matching Approach for Road Network Conflation in Different Road Patterns. ISPRS International Journal of Geo-Information, 8(2), 81.

Kilic B & Gülgen F (2020). Accuracy and similarity aspects in online geocoding services: A comparative evaluation for Google and Bing maps. International Journal of Engineering and Geosciences, 5(2), 109-119.

Koukoletsos T, Haklay M & Ellul C (2012). Assessing data completeness of VGI through an automated matching procedure for linear data. Transactions in GIS, 16(4), 477-498.

Lei T & Lei Z (2019). Optimal spatial data matching for conflation: A network flow-based approach. Transactions in GIS, 23(5), 1152-1176.

Li L & Goodchild M F (2011). An optimisation model for linear feature matching in geographical data conflation. International Journal of Image and Data Fusion, 2(4), 309-328.

Lynch M & Saalfeld A (1985). Conflation: automated map compilation – a video game approach. Proceedings of Autocarto 7, 11–14 March 1985 Washington, DC, USA, 343-352.

Memduhoğlu A & Başaraner M (2018). Possible contributions of spatial semantic methods and technologies to multi-representation spatial database paradigm. International Journal of Engineering and Geosciences, 3(3), 108-118.

Mustière S & Devogele T (2008). Matching networks with different levels of detail. GeoInformatica, 12(4), 435-453.

Olteanu-Raimond A M, Mustiere S & Ruas A (2015). Knowledge formalization for vector data matching using belief theory. Journal of Spatial Information Science, 2015(10), 21-46.

Pourabdollah A, Morley J, Feldman S & Jackson M (2013). Towards an authoritative OpenStreetMap: conflating OSM and OS OpenData national maps' road network. ISPRS International Journal of Geo-Information, 2(3), 704-728.

Rosen B & Saalfeld A (1985). Match criteria for automatic alignment. Proceedings of 7th international symposium on computer-assisted cartography (Auto-Carto 7), 11–14 March 1985 Washington, USA, 1-20.

Ruiz J J, Ariza F J, Urena M A & Blázquez E B (2011). Digital map conflation: a review of the process and a proposal for classification. International Journal of Geographical Information Science, 25(9), 1439-1466.

Saalfeld A (1988). Conflation automated map compilation. International Journal of Geographical Information System, 2(3), 217-228.

Samal A, Seth S & Cueto 1 K (2004). A feature-based approach to conflation of geospatial sources. International Journal of Geographical Information Science, 18(5), 459-489.

Song W, Keller J M, Haithcoat T L & Davis C H (2011). Relaxation-based point feature matching for vector map conflation. Transactions in GIS, 15(1), 43-60.

Şen A (2013). The applicability of artificial intelligence methods for the selection/elimination process to the stream networks in cartographic generalization. Doctoral Thesis, Yildiz Technical University, Istanbul, Turkey.

Volz S (2006). An iterative approach for matching multiple representations of street data. Proceedings of the ISPRS Workshop on Multiple Representation and Interoperability of Spatial Data, Hannover, Almanya, 22–24 Feb 2006, 36(Part 2/W40), 101–110.

Xavier E, Ariza-López F J & Ureña-Cámara M A (2016). A survey of measures and methods for matching geospatial vector datasets. ACM Computing Surveys (CSUR), 49(2), 39.

Xiong D & Sperling J (2004). Semiautomated matching for network database integration. ISPRS journal of photogrammetry and remote sensing, 59(1-2), 35-46.

Yang B, Luan X & Zhang Y (2014). A pattern-based approach for matching nodes in heterogeneous urban road networks. Transactions in GIS, 18(5), 718-739.

Yuan S & Tao C (1999). Development of conflation components. Proceedings of geoinformatics, 99, 1-13.

© Author(s) 2021.
This work is distributed under https://creativecommons.org/licenses/by-sa/4.0/