Sigma J Eng & Nat Sci 34 (4), 2016, 527-542

Sigma Journal of Engineering and Natural Sciences
Sigma Mühendislik ve Fen Bilimleri Dergisi

Research Article / Araştırma Makalesi

AN EXPERIMENT ON DISTANCE METRICS USED FOR ROAD MATCHING IN DATA INTEGRATION

Müslüm HACAR*, Türkay GÖKGÖZ

Yıldız Technical University, Faculty of Civil Engineering, Department of Geomatic Engineering, Esenler-İSTANBUL

Received/Geliş: 08.04.2016 Revised/Düzeltme: 31.07.2016 Accepted/Kabul: 21.11.2016

ABSTRACT

Decision makers and researchers need datasets from different sources to analyze, combine, or create new spatial datasets. The same entity may be represented with different geometries, topologies, and attributes in different datasets due to differences in production, such as projection, scale, accuracy, purpose, and date. The geometries, topologies, and attributes of objects are often used when combining and integrating the datasets from different sources. Matching spatial datasets is one of the most important phases of data integration. Many algorithms have been developed to match datasets using several parameters inspired by geometric, topological, and attribute similarities. They generally find the similarities between objects in different datasets and create relations between each object in order to analyze, combine, update, and transfer data. The differences in geometries, topologies, and attributes make the matching process difficult. The research problem is the critical selection of similarity parameters to ensure the satisfactory matching results. The scope of this paper was limited with distance metrics. In this study, it was aimed to determine the suitable distance metrics measured from point to point and from point to line, which are widely used as parameters in road matching. Two road datasets in different databases were automatically matched using these metrics by employing a plugin of an open desktop software. Automatic matching results were compared to manual matching results to determine the success of each matching process. Consequently, it was shown that none of these metrics for road matching was adequate on its own. However, the distance between centroids of roads and Hausdorff distances were more satisfactory.

Keywords: Distance metrics, matching, integration, conflation, road network.

VERİ ENTEGRASYONUNDA YOL EŞLEMELERİ İÇİN KULLANILAN MESAFE ÖLÇÜLERİ ÜZERİNE BİR DENEY

ÖZ

Karar alıcılar ve araştırmacılar analiz etmek, birleştirmek ya da yeni verisetleri oluşturmak için farklı kaynaklardan gelen verisetlerine ihtiyaç duyarlar. Aynı varlık; projeksiyon, ölçek, doğruluk, amaç ve zaman gibi üretim farklılıkları nedeniyle, farklı verisetlerinde farklı geometri, topoloji ve özniteliklerle temsil ediliyor olabilir. Nesnelerin geometri, topoloji ve öznitelikleri, verisetlerini birleştirirken ve entegre ederken sıklıkla kullanılırlar. Mekansal verisetlerini eşlemek veri entegrasyonunun en önemli aşamalarından biridir. Verisetlerini eşlemek için geometrik, topolojik ve öznitelik benzerlikleri içeren çeşitli parametreleri kullanan çok sayıda algoritma geliştirilmiştir. Bunlar genel olarak farklı verisetlerinin nesneleri arasındaki benzerlikleri bulur ve analiz etmek, birleştirmek, güncellemek, veri transferi yapmak için ilgili nesneler arasında ilişkiler kurar. Geometri, topoloji ve özniteliklerdeki farklılıklar eşleme işlemlerini zorlaştırmaktadır. Araştırma problemi, kabul edilebilir eşleme sonuçlarını elde etmek için benzerlik parametrelerinin kritik seçimidir. Makalenin kapsamı uzunluk ölçüleri ile sınırlandırılmıştır. Bu çalışmada, yol eşlemelerinde sıklıkla kullanılan, noktadan noktaya ve noktadan çizgiye ölçülen uygun mesafe ölçülerini belirlemek amaçlanmıştır. Farklı veritabanlarındaki iki yol veriseti bir açık kaynak yazılımın eklentisi ile bu ölçüler kullanılarak otomatik eşlenmiştir. Her bir eşleme işleminin başarısını belirlemek için otomatik eşleme sonuçları manuel eşleme sonuçları ile karşılaştırılmıştır. Sonuç olarak, yol eşlemeleri için bu ölçülerin hiç birinin tek başına yeterli olamayacağı görülmüştür. Ancak, ağırlık merkezleri arasındaki mesafe ve Hausdorff mesafeleri daha iyi sonuçlar vermiştir.

Anahtar Sözcükler: Uzunluk ölçüleri, eşleme, entegrasyon, bütünleştirme, yol ağı.

* Corresponding Author/Sorumlu Yazar: e-mail/e-ileti: mhacar@yildiz.edu.tr, tel: (212) 383 53 40

1. INTRODUCTION

Matching is a widely used procedure, especially in map conflation, updating and combining processes. It generally finds the similarities between objects in different datasets and creates relations between each object in order to analyze, combine, update, and transfer data. Also, it ensures the facilities about quality assesment for generalization (e.g. evaluating quality based on the distance between original and generalized lines) and analysis for travel behavior and route modelling (e.g. making positionally inaccurate real time GPS data significant and usable by matching with available road network data).

In general terms, the matching procedure can be defined as the establishment of relations between different datasets that represent the same entities. The relations can be viewed as bridges that identify and connect the objects in different datasets, thereby enabling the datasets to be interoperable. Many algorithms have been developed to match spatial data from different sources automatically. Matching algorithms are usually designed to consider the similarities between objects. Geometry, topology or attribute similarities can be the components of the same similarity equations in complex matching processes [1,2]. Moreover, the features in a dataset can be matched to the features in another dataset using one, two, or all three criteria (i.e., geometries, topologies, and attributes) according to the complexity of the datasets. Any one of these criteria can be determined in the similarity equations. Feature type is highly important in selecting a matching method. The Euclidian distance in Equation(1), which is the shortest distance between points, can be used as a threshold for point-to-point matching; meanwhile, the Hausdorff distance in Equation(2), which is the longest of the shortest distances between lines, is suitable for line-to-line matching [2,3]. The Hausdorff distance provides the spatial relationship between line features quantitatively for automated cartographic analyses. In addition to being used frequently in matching procedures for conflation processes, it is also employed to determine the amount of deviation of generalized lines from their original locations:

$$D_E = \sqrt{(X_2 - X_1)^2 + (Y_2 - Y_1)^2}, \tag{1}$$

$$D_H = max(min(d_1, d_2)). \tag{2}$$

The patterns and scales of the features should be considered in terms of geometric similarity. For example, a simple method that is adequate to match grid-patterned road networks may be inadequate to match circular or complex-patterned ones. For a building represented as a point in both datasets, Euclidian distance may be adequate for matching, while it may not be appropriate for the same building represented as an area in a larger scale dataset.

Lynch and Saalfeld [4] defined map conflation as “Combining two digital map files to produce a third map file which is better than each of source maps.” In their initial attempts to solve the matching problem, the geometric incompatibilities between datasets were considered [4, 5, and 6]. Saalfeld [7] determined that the matching procedures are more successful with geometrically aligned data. Deretsky and Rodny [8] developed a method using the standard operators of relational database management systems on road lines, intersection points, and their attributes to integrate geometric and attribute data of two digital maps.

Cobb et al. [9] developed a new method to match attribute-rich datasets in Vector Product Format developed by the National Imaging and Mapping Agency in the United States. In addition to conventional proximity, geometry, and topology, this method uses several semantics, such as data quality and feature code. This method is based on the semantic similarity of attributes and the shape similarity of lines, and it processes the data iteratively with respect to the selected criteria, starting from the strongest criterion and moving to the weaker ones until no new feature pairs are matched, like in Saalfeld’s [7] method.

Walter and Fritsch [10] matched the road networks at close scales using statistical analysis. Their method initially aligns the road networks through affine transformation. Buffers are created around all of the road lines of the first dataset, and the road lines of the second dataset in these buffers are identified as matching candidates. The candidates are compared according to their locations, shapes, directions, and topologies. The roads to be matched are determined via the statistical analysis.

Yuan and Tao [3] divided conflation into two general classes, as follows: "horizontal" (edge-matching) and "vertical" (overlapping). These researchers highlighted the matching procedure as the hardest step.

Kang [11] classified the matching procedure as "spatial" or "non-spatial." He stated that spatial matching can use geometric measures and topological features for point, line, and area objects, while non-spatial matching can use statistical or rule-based methods.

In their matching study on two datasets composed of polygons, Yuan and Tao [3] identified matches in the second dataset using the centroids in the first dataset through the raster shrink method. In this way, they performed attribute transfer between matching areas.

The method developed by Xiong and Sperling [12] matches road features semi-automatically. Point, segment, and edge matchings are performed automatically, while matching pairs are manually checked. This control step prevents mistakes from being repeated and improves performance and reliability.

Samal et al. [13] compared the attributes of polygon features and determined similarities. However, they found that features with significantly different attributes could not be matched. To measure the matching ability of features in terms of shape, the buffers were analyzed, and the important landmarks for polygon features were identified to calculate similarities between datasets. Finally, a similarity score was determined for each feature.

Zhang and Meng [14] matched the road layers in the Basis digital land cover model (DLM) and TeleAtlas using topology and the unsymmetrical buffer growing method. Following this, they integrated postal data, including building numbers, into TeleAtlas.

Mustière and Devogele [1] studied road networks with different scales. According to their proposed method, the matching candidate nodes are initially determined in a simple fashion according to the data-dependent distance threshold, and the matching candidate arcs are determined according to the Hausdorff distance. In this stage, a semi-Hausdorff distance between two arcs that is calculated from the most detailed to the least detailed one is used instead of the conventional Hausdorff distance. To determine the matched node pairs, specific measures are used, such as the network topology and the clockwise ordering of candidate arcs between the candidate nodes. The closest path approach is employed to determine the matched arcs. In this study, the success of the matching procedure could not be determined satisfactorily because both datasets came from same source (i.e., the French National Institute of Geographic and Forest Information) even though their scales were different. This method was enhanced in further studies [15,16] using belief theory and applied to spatial data from different sources.

Kim et al. [17] matched the polygon features in large-scale datasets that included buildings using Voronoi diagrams, triangulation, and geometric measurements. Identifying the landmarks from attribute data, Voronoi cells were established, and then triangulation was performed. The geometric measures of triangles (i.e., area and circumference) were compared to calculate similarities and conduct the matching procedure.

Li [18] and Li and Goodchild [19] developed an optimization model for the matching procedure. This model calculates total similarities for the features to be matched and matches the relevant features based on affine transformation parameters that maximize the similarities.

Song et al. [20] carried out the matching procedure by creating a confidence matrix. A distance threshold was determined based on the maximum shift between two datasets. Points in the second dataset that were closer to each other than the set distance threshold were considered matching candidates. The number of lines connecting to the point in the first dataset was compared to the number of lines connecting to each matching candidate point. In the confidence matrix, a value of "1" was assigned to each matching candidate point with an equal number of lines in the first dataset. After two iterations with the determined compatibility function, the points with maximum confidence values were matched.

Pourabdollah et al. [21] performed a matching procedure to increase the quality of open data. The Open Street Map data, which are constantly updated but are deficient in terms of attributes, were conflated with the attribute-rich Ordnance Survey data to obtain an up-to-date and attribute-rich dataset.

Yang et al. [22] classified the lines constituting the road network according to their patterns (i.e., atomic patterns and composite patterns) to determine matching candidates. In this method, the lines meeting the conventional geometrical and topological criteria are matched when they are in the same pattern class.

Bierlaire et. al. [23] proposed a probabilistic map matching approach to relate GPS data with road network to generate meaningful paths. In the approach, after a set of potential true paths are generated, a likelihood with each of them is determined. In order to calculate the likelihood of the data for a specific path, both GPS coordinates and temporal information (i.e. speed and time) are used.

Fan et. al. [24] proposed to match the OpenStreetMap road network with authority data. In the algorithm, first urban blocks, represented by polygons surrounded by their surrounding streets, are extracted. Then, by checking the topologies, the algorithm assigns road lines to edges of urban blocks. In the matching process, overlapping areas of blocks are used to match these polygons during the first step. In the second step, after edges of a matched polygon pair are matched with each other, road lines assigned to the same matched pair of urban block edges are matched with each other.

Kang et. al. [25] developed an algorithm to associate geometric relationships between sidewalk and street segments. The algorithm contains three parameters: the distance between streets and sidewalk segments, the angle between sidewalk and street segments, and the difference between the lengths of matched sidewalk and street segments.

In the literature there have been many studies about matching algorithms. Lots of them present ready-to-use algorithms with multiparameters. However, in this paper, matching process was carried out with only distance metrics to evaluate distance parameters. The present aimed to examine the five distance metrics in road matching (i.e., Euclidean distance between points, Euclidean distance between centroids, minimum Euclidean distance from point to line, Hausdorff distance, and semi-Hausdorff distance). An experiment was conducted using a plugin from an open desktop software program using two road datasets in different databases. The results were evaluated according to manual matching.

2. MATCHING METHODS USING DISTANCE METRICS

In this section five distance metrics were explained with mathematical equations and geometrical representations. In order to carry out the matching processes with these distances, user needs to specify a threshold *(T)* to determine certain matching pairs. Once *T* is specified by the user, the steps below are carried out for each method.

2.1. Matching Using the Distance between Points

Step 1: Lines composed of the same number of segments in different datasets are marked as the matching candidates. For example, the lines *m* and *n* in Figure 1 are composed of two segments, while line *p* comprises a single segment. Therefore, the lines *m* and *n* are marked as the matching candidates.

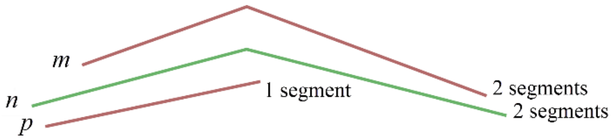


Figure 1. Matching candidates (lines *n* and *m*).

Step 2: Euclidian distances d_{mk-nl} are calculated with Equation(1) between all the points of line *m* and all the points of line *n* (Figure 2). Here, *k* and *l* are the numbers of points for lines *m* and *n*, respectively.

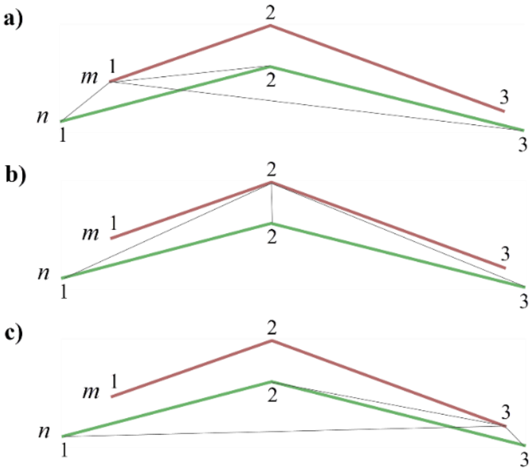


Figure 2. Euclidean distances (thin gray lines) between all points of lines *m* and *n*.

Step 3: The shortest distance at each point is determined (d_{m1-n1}, d_{m2-n2} and d_{m3-n3}; Figure 3).

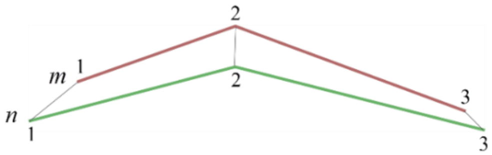


Figure 3. Shortest Euclidean distance (thin gray line) at each point.

Step 4: The maximum of the shortest Euclidian distances is determined ($d_{max} = d_{m1-n1}$) (Figure 4).

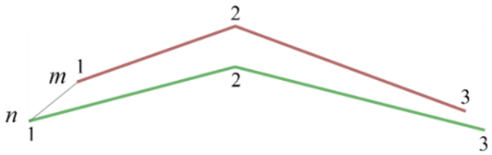


Figure 4. Maximum of the shortest Euclidian distances (d_{max}; thin gray line).

Step 5: If $d_{max} < T$, line n and line m are matched.

2.2. Matching Using the Distance between Centroids

As the scale changes, geometries of the objects also change. A centroid is the geometrical parameter that is least affected from the scale changes. Thus, a centroid may be a good measure to match data with different scales.

Step 1: Centroid coordinates of each line in the source and target datasets are calculated using the following formula:

$$X_g = \sum_{i=1}^{w} \frac{S_i x_i}{\sum S} \qquad Y_g = \sum_{i=1}^{w} \frac{S_i y_i}{\sum S}, \tag{3}$$

Here, X_g and Y_g are the centroid coordinates of a line, S_i is the length of segment *i* of the line, x_i and y_i are the midpoint coordinates of the segment *i*, *w* is the total number of segments, and *S* is the total length of the line. Figures 5a and b show the midpoints and centroids of the segments and lines, respectively.

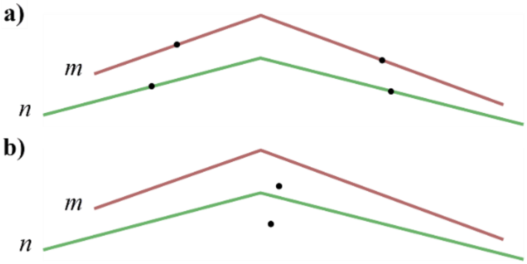


Figure 5. (a) Midpoints of segments and (b) centroids of lines (black points).

Step 2: The Euclidian distances between the centroids of two candidate lines are calculated.

Step 3: The lines with a Euclidian distance that is smaller than the threshold T are matched.

2.3. Matching Using the Minimum Distance from a Point to a Line

Step 1: The shortest distances between the points of each line in a dataset and the lines in the other dataset are calculated. Here, d_{mk-n} is the shortest distance between point k of line m and line n, while d_{nl-m} is the shortest distance between point *l* of line *n* and line *m*. For example, as shown in Figure 6, the shortest distances between line m and line n are $D_{m-n} = \{d_{m1-n}, d_{m2-n}, d_{m3-n}\}$ and $D_{n-m} = \{d_{n1-m}, d_{n2-m}, d_{n3-m}\}$. As there are no real perpendiculars from the points 1 and 3 of line *n* to line *m*, the lines connecting points 1 and 3 of line *n* to points 1 and 3 of line *m*, respectively, represent the shortest distances between these points. However, it is possible to draw perpendicular lines from points 1 and 3 of line *m* to line *n*. Thus, the shortest distances at points 1 and 3 of line *m* are the lengths of the perpendiculars from these points to line *n*. Similarly, while it is possible to draw a perpendicular from point 2 of line *n* to line *m*, it is not possible from point 2 of line *m* to line *n*. Thus, while the shortest distance at point 2 of line *n* is the perpendicular distance, the shortest distance at point 2 of line *m* is the length of the line connecting point 2 of line *n* to point 2 of line *m*.

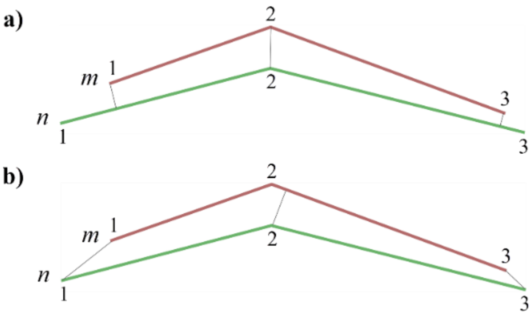


Figure 6. The shortest distances between points and lines (thin gray line).

Step 2: The minimum of the shortest distances ($d_{min} = min(D_{m-n}, D_{n-m})$ is determined. For example, as shown in Figure 7, the shortest distance calculated at point 3 of line m (d_{m3-n}) is determined as the minimum of the shortest distances.

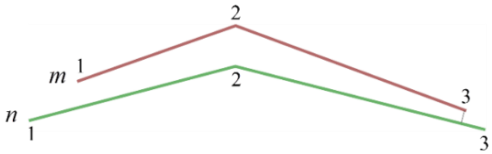


Figure 7. The minimum of the shortest distances (thin gray line).

Step 3: If $d_{min} < T$, the lines are matched.

2.4. Matching Using the Hausdorff Distance

Step 1: The shortest distances between the points of each line in a dataset and the lines in the other dataset are determined as in the "matching using the minimum distance from a point to a line" method.

Step 2: The maximum of the shortest distances ($d_{max} = max(D_{m-n}, D_{n-m})$) is determined. For example, as shown in Figure 8, the maximum of the shortest distances between lines m and n is the distance between point 1 of line n and point 1 of line m (d_{n1-m}).

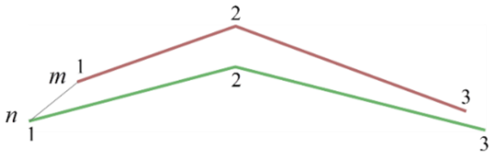


Figure 8. Maximum of the shortest distances (thin gray line).

Step 3: If $d_{max} < T$, the lines are matched.

2.5. Matching Using the Semi-Hausdorff Distance

This is the matching procedure based on the one-way Hausdorff distance from the line in the source dataset to the line in the target dataset.

Step 1: The shortest distances between the points of each line in the source dataset and the lines in the target dataset are determined. For example, as shown in Figures 6a and 6b, the shortest distances are determined from line n to m and from line m to n, respectively.

Step 2: The maximum of the shortest distances ($max(D_{m-n})$ or $max(D_{n-m})$) is determined. For example, while $max(D_{n-m}) = d_{n1-m}$ is the semi-Hausdorff distance from line n to line m (i.e., line m in the target and line n in the source datasets; Figure 8), $max(D_{m-n}) = d_{m2-n}$ is the semi-Hausdorff distance from line m to line n (i.e., line n in the target and line m in the source datasets; Figure 9).

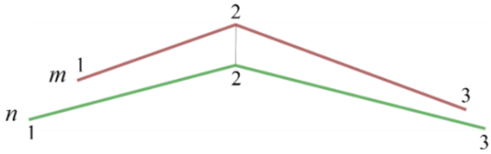


Figure 9. The semi-Hausdorff distance from line m to line n (thin gray line).

Step 3: If $max(D_{m-n}) < T$ or $max(D_{n-m}) < T$, the lines are matched.

3. EXPERIMENT AND RESULTS

An experiment was conducted using the methods mentioned above with two road datasets in different databases, and the results were compared to manual matching results assumed to be the expected results to determine the numbers of correct and incorrect matchings. The methods were automatically implemented using MatchingPlugin (MatchingPlugIn0.7.2) from the OpenJump desktop software [26].

Two datasets composed of several road centerlines in a part of the Beykoz district (10 km x 6 km) of Istanbul, Turkey were used (Figure 10). The datasets were produced by the Istanbul Metropolitan Municipality (IMM) Directorate of Geographical Information Systems and Başarsoft Information Technologies Inc. (Figure 11). The study area was further divided into three zones to facilitate the verification of automatic matching (Figure 12). Hereafter, the IMM data are referred to as the "target dataset," and the Başarsoft data are labelled the "source dataset."

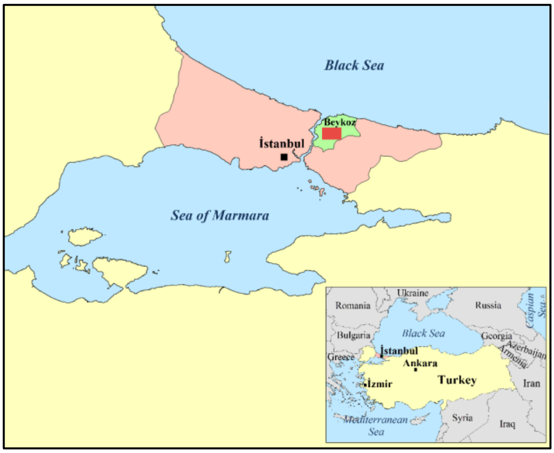


Figure 10. The study area (red rectangle) and inset map (Map data was taken from GADM database [27]).

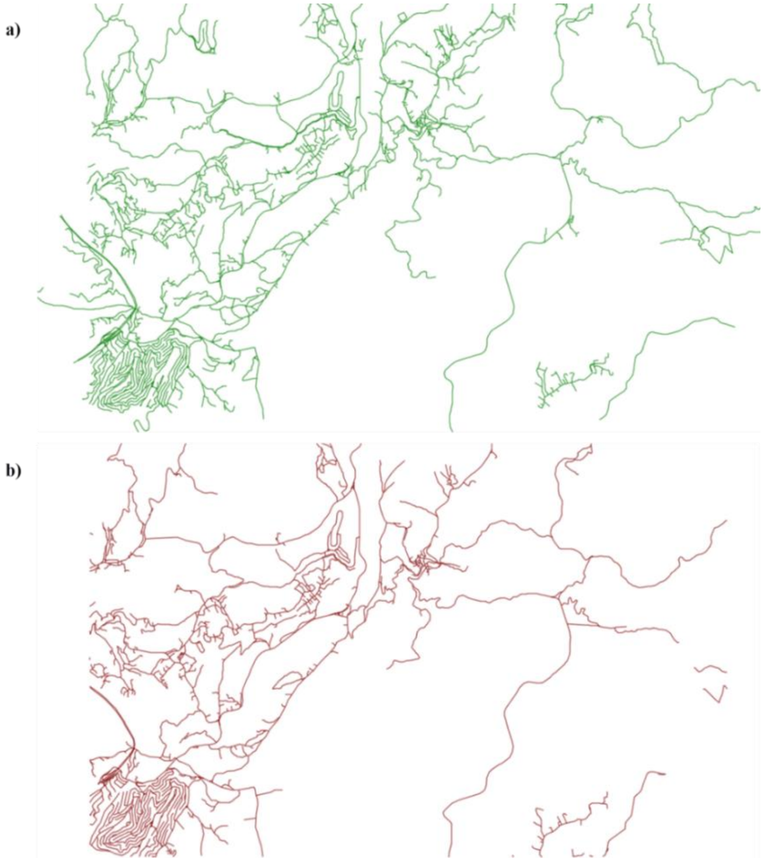


Figure 11. (a) IMM roads and (b) Başarsoft roads.

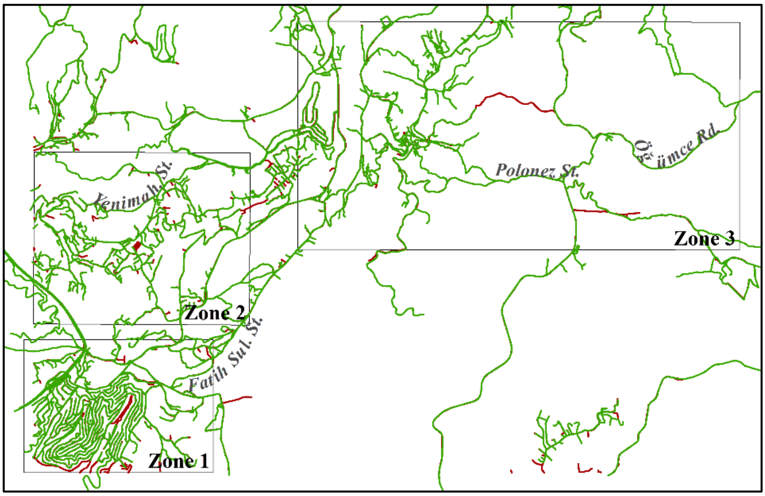


Figure 12. Both IMM (target, green) and Başarsoft roads (source, red) and zones.

While the target dataset is used in the Geographic Information System (GIS) infrastructure by IMM at scale of 1:1,000, the source dataset at scale of 1:5000 is used in navigation applications by Başarsoft. The datum of the target dataset is ITRF2005.0, and the datum of the source dataset is WGS84. There are locational and topological differences between the datasets due to the production method, date, source, scale, and so on (Figure 13). Some road names in the target dataset are not up-to-date. While the source dataset is attribute-rich, it is not geometrically up-to-date (some roads do not exist), and some roads are represented with simpler geometry.

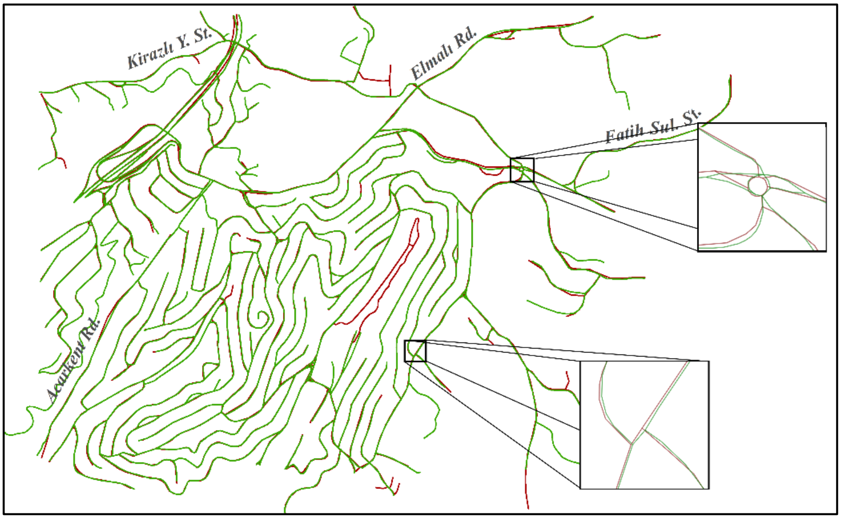


Figure 13. Overlapping view of the target (green) and source (red) datasets, and the samples showing the roads are not exactly superposed.

While some roads are represented with two or more line features in the target dataset, they are represented with a single line feature in the source dataset (or vice versa). Thus, the number of matched features in manual matching was determined by considering 1:1, 1:N, and N:1 matching conditions (Table 1). To measure the success of the automated methods with respect to the manual matching, the numbers of “Matching,” “Correct Matching,” “Incorrect Matching,” “Incomplete Matching,” and “No Matching” features have been determined.

Table 1. Numbers of manually matched features in each zone.

	Zone 1		**Zone 2**		**Zone 3**	
	Target	**Source**	**Target**	**Source**	**Target**	**Source**
Number of features	365	367	293	272	306	190
Number of matched features	332	332	248	232	215	176
Number of unmatched features	33	35	45	40	91	14
Number of matching pairs	332		248		215	

In the experiment, 1:1 and 1:N matchings were tested for all methods by specifying the distance threshold as T=10 meters. The threshold value was determined by manually measuring the maximum distances between lines in the datasets. The results obtained in Zones 1, 2, and 3 are presented as manual, automated, correct, incorrect, incomplete, and unmatched feature counts in Tables 2, 3, and 4, respectively.

Table 2. Zone 1: 1:N and 1:1 matching results.

Zone 1	**Distance between points**		**Distance between centroids**		**Minimum distance from point to line**		**Hausdorff distance**		**Semi-Hausdorff distance**	
	1:N	**1:1**	**1:N**	**1:1**	**1:N**	**1:1**	**1:N**	**1:1**	**1:N**	**1:1**
Manual	332		332		332		332		332	
Automated	35	33	238	222	332	292	185	185	326	247
Correct	24	24	157	205	17	150	148	171	44	203
Incorrect	9	7	78	7	314	121	32	7	281	13
Incomplete	2	2	3	10	1	21	5	7	1	31
Unmatched	297	299	94	110	0	40	147	147	6	85

Table 3. Zone 2: 1:N and 1:1 matching results.

Zone 2	**Distance between points**		**Distance between centroids**		**Minimum distance from point to line**		**Hausdorff distance**		**Semi-Hausdorff distance**	
	1:N	**1:1**	**1:N**	**1:1**	**1:N**	**1:1**	**1:N**	**1:1**	**1:N**	**1:1**
Manual	248		248		248		248		248	
Automated	11	10	163	151	248	208	140	139	246	176
Correct	11	10	145	149	22	126	139	138	45	154
Incorrect	0	0	17	1	226	67	0	0	201	2
Incomplete	0	0	1	1	0	15	1	1	0	20
Unmatched	237	238	85	97	0	40	108	109	2	72

Table 4. Zone 3: 1:N and 1:1 matching results.

Zone 3	**Distance between points**		**Distance between centroids**		**Minimum distance from point to line**		**Hausdorff distance**		**Semi-Hausdorff distance**	
	1:N	**1:1**	**1:N**	**1:1**	**1:N**	**1:1**	**1:N**	**1:1**	**1:N**	**1:1**
Manual	215		215		215		215		215	
Automated	12	12	115	112	214	143	96	96	205	125
Correct	12	12	98	107	36	66	93	93	58	109
Incorrect	0	0	14	1	178	69	0	0	145	2
Incomplete	0	0	3	4	0	8	3	3	2	14
Unmatched	203	203	100	103	1	72	119	119	10	90

To determine the success of the matching methods, percentages of automated, correct, and incorrect matchings were calculated (Tables 5, 6, and 7). While 99.8% of the features were automatically matched with the method that used the minimum distance from a point to a line, which represented the highest matching percentage for 1:N matching (Table 5), it was found that most of the matchings were incorrect (Table 7). However, 58.0% of the features on average were correctly matched using the semi-Hausdorff distance method, which gave the best results for 1:1

matching (Table 6). Matching using the Hausdorff distance produced the least incorrect matchings. Moreover, the method using the distance between centroids produced almost the same results as the method using semi-Hausdorff distance for 1:1 matching. The method using the distance between centroids produced the next best results after the Hausdorff distance method in terms of incorrect matching (Table 7).

Table 5. Percentages of automated matching ($(Automated/Manual) \times 100$).

	Zone 1		**Zone 2**		**Zone 3**		**Average**	
	1:N	**1:1**	**1:N**	**1:1**	**1:N**	**1:1**	**1:N**	**1:1**
Distance between points	10.5	9.9	4.4	4.0	5.58	5.58	6.85	6.52
Distance between centroids	71.7	66.9	65.7	60.9	53.5	52.1	63.6	60.0
Minimum distance from point to line	100	88.0	100	83.9	99.5	66.5	99.8	79.4
Hausdorff distance	55.7	55.7	56.5	56.1	44.7	44.7	52.2	52.1
Semi-Hausdorff distance	98.2	74.4	99.2	71.0	95.4	58.1	97.6	67.8

Table 6. Percentages of correct matching ($(Correct/Manual) \times 100$).

	Zone 1		**Zone 2**		**Zone 3**		**Average**	
	1:N	**1:1**	**1:N**	**1:1**	**1:N**	**1:1**	**1:N**	**1:1**
Distance between points	7.2	7.2	4.4	4.0	5.6	5.6	5.8	5.6
Distance between centroids	47.3	61.8	58.5	60.1	45.6	49.8	50.5	57.2
Minimum distance from point to line	5.1	45.2	8.9	50.8	16.7	30.7	10.3	42.2
Hausdorff distance	44.6	51.5	56.1	55.7	43.3	43.3	48.0	50.1
Semi-Hausdorff distance	13.3	61.1	18.2	62.1	27.0	50.7	19.5	58.0

Table 7. Percentages of incorrect matching ($(Incorrect/Manual) \times 100$).

	Zone 1		**Zone 2**		**Zone 3**		**Average**	
	1:N	**1:1**	**1:N**	**1:1**	**1:N**	**1:1**	**1:N**	**1:1**
Distance between points	25.7	21.2	0.0	0.0	0.0	0.0	8.57	7.07
Distance between centroids	32.8	3.2	10.4	0.7	12.2	0.9	18.5	1.6
Minimum distance from point to line	94.6	41.4	91.1	32.2	83.2	48.3	89.6	40.6
Hausdorff distance	17.3	3.8	0.0	0.0	0.0	0.0	5.8	1.3
Semi-Hausdorff distance	86.2	5.3	81.7	1.1	70.7	1.6	79.6	2.7

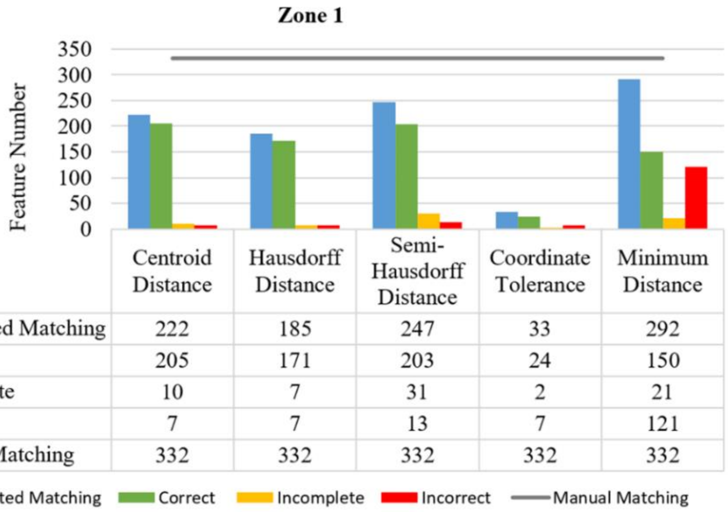
3.1. Evaluation of the Results

In road-matching tasks, evaluation of the results depends on the purpose of matching. For example, vehicles like ambulances, firetrucks, mail trucks, school buses, and taxis use navigation systems that visualize road networks. These systems must always be up-to-date and reliable. In this regard, the accuracy of the matching method should reach a level that satisfies the requirements of the navigation system [28]. In this study, the methods were evaluated according to the correct and incorrect matchings.

The results obtained in each zone with the methods using Hausdorff and semi-Hausdorff distances showed that these approaches cannot be used on their own. However, the fact that the number of incorrect matchings of these methods (especially using Hausdorff distance) was smaller than that of the others explains why these are used as one of the primary measures in most matching algorithms (Tables 2, 3, and 4).

The method using distance between centroids is not a widely used method for line matching. However, it was the second best method in terms of incorrect matching, following the Hausdorff distance method. In addition, although this method was less successful than the method using semi-Hausdorff distance in terms of correct matchings, it was more successful than the method using Hausdorff distance (Tables 2, 3, and 4). In this context, while the distance between two centroids is not adequate as a measure for use in a matching method, it is a measure that can be employed in the same way as the Hausdorff distance in algorithms containing similarity equations with different criteria.

The matching results obtained in Zone 1 are represented as a graph in Figure 14. It is obvious that the methods using distance between centroids and Hausdorff and semi-Hausdorff distances were more successful than the other approaches in terms of correct and incorrect matchings.



	Centroid Distance	Hausdorff Distance	Semi-Hausdorff Distance	Coordinate Tolerance	Minimum Distance
🟦 Automated Matching	222	185	247	33	292
🟩 Correct	205	171	203	24	150
🟨 Incomplete	10	7	31	2	21
🟥 Incorrect	7	7	13	7	121
— Manual Matching	332	332	332	332	332

Figure 14. Zone 1: 1:1 matching results.

As shown in Figure 15, all roads in Zone 1 that were incorrectly matched by all methods were positioned at the junctions. For example, a rectangular junction in the source dataset was represented as a triangular junction in the target dataset (Figure 16).

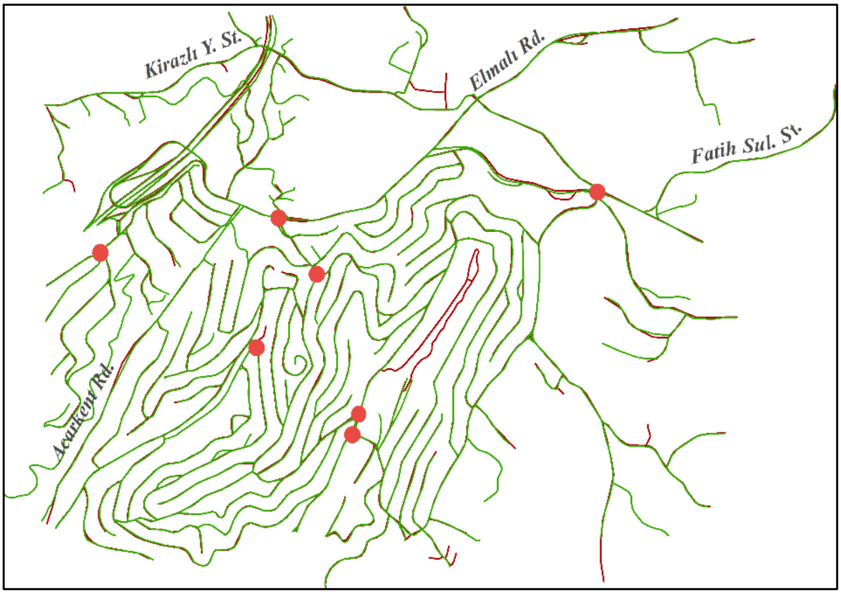
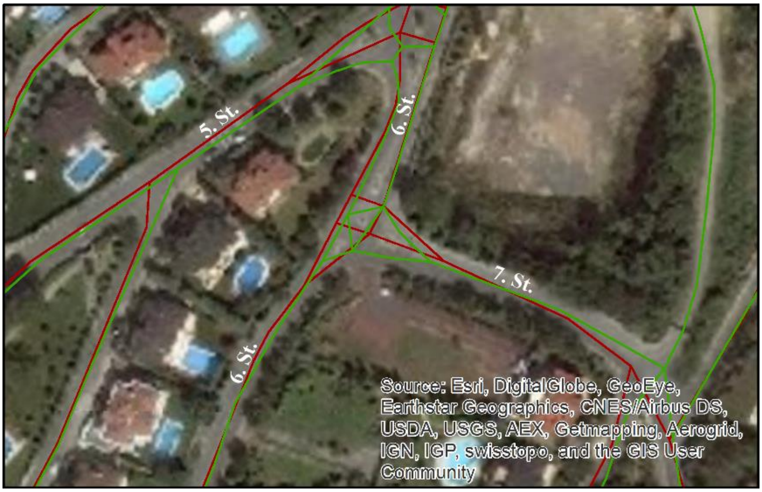


Figure 15. Positions of roads (red points) incorrectly matched by all methods in Zone 1.



Source: Esri, DigitalGlobe, GeoEye, Earthstar Geographics, CNES/Airbus DS, USDA, USGS, AEX, Getmapping, Aerogrid, IGN, IGP, swisstopo, and the GIS User Community

Figure 16. Different representations of a junction in the source (red lines) and target (green lines) datasets overlapped with satellite image [29].

4. CONCLUSION AND DISCUSSION

Each method used in this study has its own characteristics. While an object pair is determined as matched in one of the five methods, it may be unmatched in another. Also, while a distance metric is relatively accurate for matching the datasets in one region, it may fail in other regions. This means that the distance metrics are data-dependent. Hausdorff and semi-Hausdorff distances are more satisfactory that are already used in many matching methods. In the literature, many of the researchers have prefered these distance metrics since the determination of the difference between two line is considered as from points of first line to the other line. However, the distance between the centroids of lines (from point to point) seems like a useful metric for road matching as well. It was the second best method in terms of incorrect matching, following the Hausdorff distance method. Also, none of the tested methods were completely satisfactory for road matching since the purpose of this study is only the examination of the distance metrics. The similarity equations determine the success of the matching processes directly. The study also shows all of the examined methods failed at junctions.

This study was only implemented in datasets with similar large scales (1:1.000 and 1:5.000). To make a comparison between sources with multi-scale, it may be tested in datasets with middle scales (e.g. 1:25.000 and 1:100.000). Angle metrics may be examined to determine angle parameter. Different parameters such as angle, topology or attribute informations may be used together to create similarities between objects of datasets. Using the patterns of junctions in road matching may be a topic worth examining.

Acknowledgments / Teşekkür

Authors would like to thank Michaël Michaud (one of OpenJump project leaders) for his support, and IMM Directorate of Geographical Information Systems and Başarsoft Information Technologies Inc. for supplying road datasets. Authors are also grateful to the anonymous reviewers whose comments helped to improve the paper.

REFERENCES / KAYNAKLAR

[1] Mustière, S. and Devogele, T., (2008). "Matching Networks with Different Levels of Detail", GeoInformatica, 12(4): 435-453.

[2] Zhang, M., (2009). Methods and Implementations of Road-Network Matching. Ph. D. Thesis, Technical University of Munich, Münih.

[3] Yuan, S. and Tao, C., (1999). "Development of Conflation Components", The Proceedings of Geoinformatics'99 Conference, 19-21 June 1999, Ann Arbor, 1-13.

[4] Lynch, M.P. and Saalfeld, A., (1985). "Conflation: Automated Map Compilation—A Video Game Approach", Auto-Carto VII, 11-14 March 1985, Washington, D.C.

[5] Rosen, B. and Saalfeld, A., (1985). "Match Criteria for Automatic Alignment" Auto-Carto VII, 11-14 March 1985, Washington, D.C.

[6] Lupien, A. E. and Moreland, W.H., (1987). "A General Approach to Map Conflation", Auto-Carto VIII, 29 Mart- 3 April 1987, Baltimore.

[7] Saalfeld, A., (1988). "Conflation: Automated Map Compilation", International Journal of Geographical Information System, 2(3): 217-228.

[8] Deretsky, Z. and Rdony, U., (1993). "Automatic Conflation of Digital Maps", In Vehicle Navigation and Information Systems Conference, October 1993, Ottawa.

[9] Cobb, M.A., Chung, M.J., Foley III, H., Petry, F.E., Shaw, K.B. and Miller, H.V., (1998). "A Rule-Based Approach for the Conflation of Attributed Vector Data", GeoInformatica, 2(1): 7-35.

[10] Walter, V. and Fritsch, D., (1999). "Matching Spatial Data Sets: A Statistical Approach", International Journal of Geographical Information Science, 13(5): 445-473.

[11] Kang, H., (2002). Analytical Conflation of Spatial Data from Municipal and Federal Government Agencies, Ph. D. Thesis, the Ohio State University, Ohio.

[12] Xiong, D. and Sperling, J., (2004). "Semiautomated Matching for Network Database Integration", ISPRS Journal of Photogrammetry and Remote Sensing, 59(1): 35-46.

[13] Samal, A., Seth, S. and Cueto, K., (2004). "A Feature-Based Approach to Conflation of Geospatial Sources", International Journal of Geographical Information Science, 18(5): 459-489.

[14] Zhang, M. and Meng, L., (2007). "An Iterative Road-Matching Approach for The Integration of Postal Data", Computers, Environment and Urban Systems, 31(5): 597-615.

[15] Olteanu-Raimond, A.M. and Mustière, S., (2008)."Data Matching—A Matter of Belief", Proceedings of the International Symposium on Spatial Data Handling, 23–25 June 2008, Montpellier, 501-19.

[16] Olteanu-Raimond, A.M., Mustière, S. and Ruas, A., (2015). "Knowledge Formalisation for Vector Data Matching Using Belief Theory", Journal of Spatial Information Science, 10: 21-46.

[17] Kim, J.O., Yu, K., Heo, J. and Lee, W.H., (2010). "A New Method for Matching Objects in Two Different Geospatial Datasets Based on the Geographic Context", Computers & Geosciences, 36(9): 1115-1122.

[18] Li, L., (2010). Design of A Conceptual Framework and Approaches for Geo-Object Data Conflation, Ph. D. Thesis, University of California, Santa Barbara.

[19] Li, L. and 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.

[20] Song, W., Keller, J.M., Haithcoat, T.L. and Davis, C.H., (2011). "Relaxation-Based Point Feature Matching for Vector Map Conflation", Transactions in GIS, 15(1): 43-60.

[21] Pourabdollah, A., Morley, J., Feldman, S. and 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.

[22] Yang, W., Lee, D. and Ahmed, N., (2014). "Pattern Based Feature Matching for Geospatial Data Conflation", GEOProcessing, March 2014, Barcelona.

[23] Bierlaire, M., Chen, J., and Newman, J., (2013). "A probabilistic map matching method for smartphone GPS data", Transportation Research Part C: Emerging Technologies, 26: 78-98.

[24] Fan, H., Yang, B., Zipf, A., and 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.

[25] Kang, B., Scully, J. Y., Stewart, O., Hurvitz, P. M., and Moudon, A. V., (2015). "Split-Match-Aggregate (SMA) algorithm: integrating sidewalk data with transportation network data in GIS", International Journal of Geographical Information Science, 29(3): 440-453.

[26] Michaud M., MatchingPlugIn Tutorial for Version 0.7.2., [Internet] http://sourceforge.net/projects/jump-pilot/files/OpenJUMP_plugins/More%20Plugins/Matching%20PlugIn/MatchingPlugIn0.7.2.pdf/download, [Accessed on 15.03.2016].

[27] Global Administrative Areas, GADM database, California, USA, [Internet] http://www.gadm.org/, [Accessed on 10.08.2016]

[28] Hacar, M., (2015). Mekânsal Veri Altyapilarinda Geometrik Entegrasyon, M. Sc. Thesis, Yıldız Technical University, İstanbul.

[29] Esri, DigitalGlobe, GeoEye, Earthstar Geographics, CNES/Airbus DS, USDA, USGS, AEX, Getmapping, Aerogrid, IGN, IGP, swisstopo, and the GIS User Community.