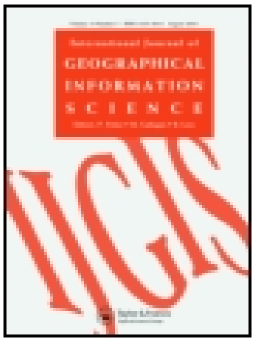

International Journal of Geographical Information Science

ISSN: 1365-8816 (Print) 1362-3087 (Online) Journal homepage: http://www.tandfonline.com/loi/tgis20

A simplified linear feature matching method using decision tree analysis, weighted linear directional mean, and topological relationships

Ick-Hoi Kim, Chen-Chieh Feng & Yi-Chen Wang

To cite this article: Ick-Hoi Kim, Chen-Chieh Feng & Yi-Chen Wang (2016): A simplified linear feature matching method using decision tree analysis, weighted linear directional mean, and topological relationships, International Journal of Geographical Information Science, DOI: 10.1080/13658816.2016.1267736

To link to this article: http://dx.doi.org/10.1080/13658816.2016.1267736

View supplementary material

Published online: 26 Dec 2016.

Submit your article to this journal

View related articles

View Crossmark data

Taylor & Francis
Taylor & Francis Group

A simplified linear feature matching method using decision tree analysis, weighted linear directional mean, and topological relationships

Ick-Hoi Kim, Chen-Chieh Feng and Yi-Chen Wang

Department of Geography, National University of Singapore, Singapore

ABSTRACT

Linear feature matching is one of the crucial components for data conflation that sees its usefulness in updating existing data through the integration of newer data and in evaluating data accuracy. This article presents a simplified linear feature matching method to conflate historical and current road data. To measure the similarity, the shorter line median Hausdorff distance (SMHD), the absolute value of cosine similarity (aCS) of the weighted linear directional mean values, and topological relationships are adopted. The decision tree analysis is employed to derive thresholds for the SMHD and the aCS. To demonstrate the usefulness of the simple linear feature matching method, four models with incremental configurations are designed and tested: (1) Model 1: one-to-one matching based on the SMHD; (2) Model 2: matching with only the SMHD threshold; (3) Model 3: matching with the SMHD and the aCS thresholds; and (4) Model 4: matching with the SMHD, the aCS, and topological relationships. These experiments suggest that Model 2, which considers only distance, does not provide stable results, while Models 3 and 4, which consider direction and topological relationships, produce stable results with levels of accuracy around 90% and 95%, respectively. The results suggest that the proposed method is simple yet robust for linear feature matching.

ARTICLE HISTORY
Received 10 September 2015
Accepted 29 November 2016

KEYWORDS
Feature matching; data conflation; Hausdorff distance; cosine similarity; historical data

1. Introduction

Advances in location-based technologies have provided us with more accurate data sets. For instance, in 2002 the MAF/TIGER Enhancement Program has been initiated to improve the positional accuracy of the U.S. census spatial data (Trainor 2003). Additionally, a variety of data sources are available through the Internet, which are often conflated to acquire more knowledge than using a single data source alone (Li and Goodchild 2011). Most data conflation studies have focused on feature matching of heterogeneous but contemporary data sets, specifically on approaches to update existing data through newer data and to evaluate data accuracy (Saalfeld 1988, Van Niel and McVicar 2002). At the conceptual level, these general approaches are equally applicable to historical data as they are for recent data. This study aims to implement a new data conflation method to compare historical road data and recent road data.

CONTACT Chen-Chieh Feng geofcc@nus.edu.sg
Supplemental data for this article can be accessed here.

Conflation of historical data, however, faces at least three challenges. First, geometrical features from historical data sources (e.g., topographic paper maps) are often more generalized and simplified. As such, positional errors are inevitable when digitizing the features from the maps. Second, historical data have lower levels of detail (LoDs) than recent data. For example, historical maps mostly employ single lines to represent road networks, while recent maps may use multiple lines to represent expressways. Third, attributes, such as road and building names, are often lacking from historical maps. As Safra *et al.* (2010) argued, location only feature matching without considering attributes is not easy. Despite these challenges, accurate conflation of historical data is necessary for spatial-temporal analysis.

Numerous studies have focused on feature matching based on geometrical similarity (Walter and Fritsch 1999, Beeri *et al.* 2004, Samal *et al.* 2004, Zhang and Meng 2007, Hastings 2008, Li and Goodchild 2011), topological relationships (Zhang and Meng 2007, Hope and Kealy 2008, Yang *et al.* 2014), and semantic similarity (Samal *et al.* 2004, Du *et al.* 2012). To improve the matching accuracy, multiple similarity measures have been employed based on combinations of distances between features, angle or direction differences, length differences, topological relationships, and attributes (Samal *et al.* 2004, Zhang and Meng 2007, Li and Goodchild 2011). However, methods employing multiple properties can lead to higher level of uncertainty and subjectivity. For instance, the matching methods considering topological relationships of complicated junctions require complicated rules to be identified by experts (Zhang and Meng 2007). Likewise, approaches involving geometrical similarity commonly require user-defined weights and thresholds (Beeri *et al.* 2004, Samal *et al.* 2004, Zhang and Meng 2007, Li and Goodchild 2011). Although iteration methods have been proposed to automatically obtain the thresholds (Yang *et al.* 2013, Tong *et al.* 2014), additional parameter values are needed to stop the iterations, and it is uncertain when the iterations terminate.

Accordingly, it is desirable to utilize as few similarity measures as possible. The method by Tong *et al.* (2014) is noticeable because they used only one similarity measure modified from the Hausdorff distance (HD). Following their method, we conducted a pilot study using our study area in Singapore but found the results unstable, with accuracies ranging between 65% and 82%. We therefore examine the limitations of prior work, and propose a simple but robust method considering distance, angle, and topological relationship. To compare the performances of different components, multiple models with different configurations of distance, angle, and topological relationship are designed: (1) Model 1: one-to-one matching based on distance; (2) Model 2: matching with only distance thresholds; (3) Model 3: matching with distance and angle thresholds; and (4) Model 4: matching with distance, angle, and topological relationship. Each model is improved based on the outputs of the previous models.

The novelty of our method is threefold. First, instead of using the direction between a start point and an end point, the weighted linear directional mean (WLDM) calculating directions weighted by line segments is employed. Then, the difference of the WLDM is converted to the absolute value of cosine similarity (aCS) to measure the angle between matched features. Second, to automatically derive thresholds for distance and angle, decision tree analysis is adopted, and no iteration is needed as opposed to Yang *et al.* (2013) and Tong *et al.* (2014). Third, for more sophisticated road network scenarios, simple topological relationships are employed with no prior knowledge but one simple rule to check the directions of roads from matched junctions.

The remainder of this article is organized as follows. Section 2 introduces data sets, workflow, and models with different configurations. Section 3 presents the results of the different models, and discusses the model performance and accuracy. Section 4 presents conclusions.

2. Materials and methods

Since the main application of this study is to compare historical with recent data, two data sets that are 40 years apart are prepared (Table 1). One is a road network data set digitized manually from a 1974 historical paper map. In Singapore, the largest scale of the paper map accessible to the public is 1:50,000. The road networks digitized from the paper map have higher positional uncertainties due to the map scale and lack of projection metadata, hampering the feature matching process. The other data set is from the Singapore authoritative land-based data, named the Singapore Street Map. This recent data set is more accurate than the historical data set, and has a higher LoD, representing expressways as multilanes. Hereafter, we term the digitized historical data set the target data set, and the Singapore authoritative data set the reference data set.

Many linear feature matching methods are initialized by collecting matching candidates with buffer analysis (Zhang and Meng 2007, Seo and O'Hara 2009, Yang *et al*. 2014). Hence, for data preprocessing, buffering is conducted for both data sets (Figure 1). The buffer distance used is 25 m, a standard error tolerance used for 1:50,000 scale maps (Van Niel and McVicar 2002), the map scale of the paper map in this study. After buffer analysis, the buffer polygons of the reference data set are intersected with the linear features of the target data set, and the intersected linear features are created with the linear feature identification numbers of both data sets. The outputs are then dissolved based on the unique combination of the identification numbers of the target data set and the reference data set. Road junctions are also created by intersect analysis from each road network data set. Each junction has the linear feature identification numbers linked to the junction.

2.1. *Model 1: the shorter line median Hausdorff distance (SMHD)*

Model 1 considers only one-to-one matching based on the shorter line median Hausdorff distance (SMHD) proposed by Tong *et al*. (2014). The SMHD is a variant of the HD that measures the distance between two lines based on their vertices:

$$\begin{aligned} \mathrm{HD}(\mathrm{P},\mathrm{Q}) &= max\{h(P,Q),h(Q,P)\} \\ \mathrm{h}(\mathrm{P},\mathrm{Q}) &= \max_{p\in P}\min_{q\in Q}\|p-q\|, \end{aligned} \tag{1}$$

Table 1. Data sets for linear feature matching.

	Historical data (Target data set)	Singapore authoritative data (Reference data set)
Year	1974	2014
Projection	Malayan Rectified Skew Orthomorphic	SVY21 Transverse Mercator
Scale	1:50,000	N.A.

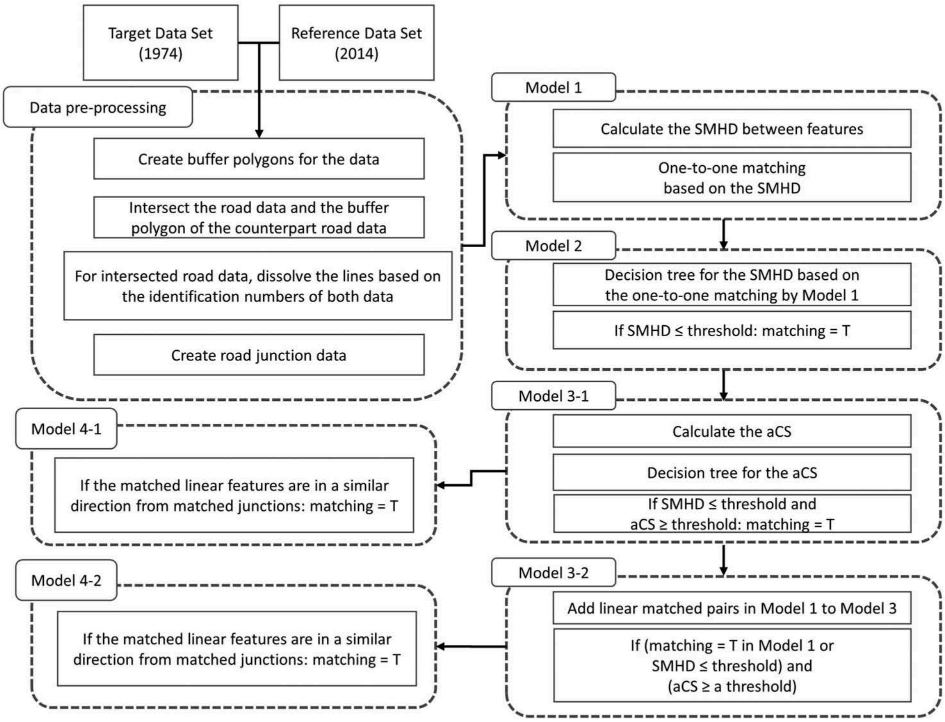
Target Data Set (1974)
Reference Data Set (2014)
Data pre-processing
Create buffer polygons for the data
Intersect the road data and the buffer polygon of the counterpart road data
For intersected road data, dissolve the lines based on the identification numbers of both data
Create road junction data
Model 1
Calculate the SMHD between features
One-to-one matching based on the SMHD
Model 2
Decision tree for the SMHD based on the one-to-one matching by Model 1
If SMHD ≤ threshold: matching = T
Model 3-1
Calculate the aCS
Decision tree for the aCS
If SMHD ≤ threshold and aCS ≥ threshold: matching = T
Model 4-1
If the matched linear features are in a similar direction from matched junctions: matching = T
Model 3-2
Add linear matched pairs in Model 1 to Model 3
If (matching = T in Model 1 or SMHD ≤ threshold) and (aCS ≥ a threshold)
Model 4-2
If the matched linear features are in a similar direction from matched junctions: matching = T

Figure 1. Preprocessing and models of linear feature matching in the proposed method. Matching = T means the final matching pairs.

where $\mathrm{HD}(\mathrm{P},\mathrm{Q})$ is the HD between the line P and the line Q, and $\mathrm{h}(\mathrm{P},\mathrm{Q})$ is the longest of the shortest distances from every vertex from the point set $\{p_1, p_2, \ldots, p_m\}$ of P to all vertices from the point set $\{q_1, q_2, \ldots, q_n\}$ of Q.

While the HD has been widely used to measure the shape similarity between features (Yuan and Tao 1999, Li and Goodchild 2011, Yang *et al.* 2013), it is not suited to one-to-many matching. For instance, when the line q_1 in Figure 2(a) is split into q_{1a} and q_{1b}, the HDs in Figure 2(b) are longer than the HD in Figure 2(a). If a threshold is smaller than the HDs, the line will be classified as incorrectly matched pairs. The SMHD addresses this issue by calculating the HD from a shorter line to a longer line.

In the SMHD, the distance between two lines is not the vertex-to-vertex distance. Instead, it is the perpendicular distance from a vertex on the shorter line to the longer line. Only in rare cases when such distance does not exist because no perpendicular line

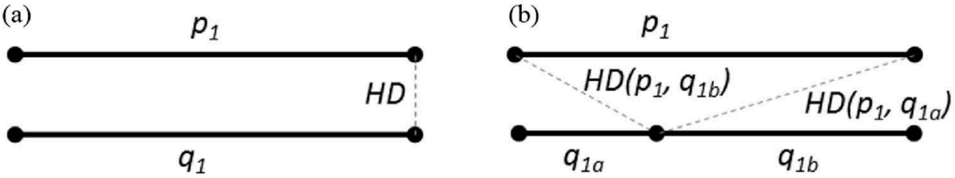

Figure 2. The example of the HD between lines p_1 and $q_{1.}$

connecting the vertex on the shorter line to the longer line is found, the distance is reverted back to vertex-to-vertex distance between the vertex on the shorter line and the vertices of the longer line segment. The SMHD is calculated as:

$$\mathrm{SMHD}(\mathrm{P},\mathrm{Q}) = \left\{ \begin{array}{l} \underset{p\in P}{\mathrm{median}}\,\underset{ql\in Q}{\min}\|p - ql\|, if\ length(P) < length(Q) \\ \underset{q\in Q}{\mathrm{median}}\,\underset{pl\in P}{\min}\|q - pl\|, if\ length(Q) < length(P) \end{array} \right\}, \tag{2}$$

where *pl* and *ql* are line segments of the lines P and Q, respectively. The SMHD is the HD selected not by maximum value but by median value from a shorter line to a longer line.

The SMHD is very simple compared with other linear feature matching approaches that combine line length, arc, chord, angle and direction, vertex matching, or topological relationships. Nevertheless, Model 1 is limited to one-to-one matching. Additional processing is needed to capture many-to-many matching.

2.2. *Model 2: decision tree analysis*

Model 2 is a threshold-based matching model that handles many-to-many matching. To determine the threshold to classify correctly and incorrectly matched pairs, decision tree analysis, a classification method based on the information theory (Quinlan 1986) is used. Specifically, the C4.5 algorithm (Quinlan 1993) is adopted to generate decision trees (Supplementary 1). The one-to-one matching results generated by Model 1 are used as pseudo-observations and a training data set.

To split the input data, decision tree analysis employs the entropy suggested by Shannon (1948), and selects the attributes or the values of an attribute deriving the maximum value in information gain. The ID3 algorithm in C4.5 employs a greedy search to obtain a threshold to split a continuous variable. Once the threshold is determined by the decision tree analysis, the matched pairs are classified using the rules in Equation (3). The advantages of the decision tree analysis are its simplicity and nondependency on iteration.

$$Matching = \left\{ \begin{array}{c} T, if\ \ SMHD \le Threshold \\ F, otherwise \end{array} \right\}. \tag{3}$$

2.3. *Model 3: weighted linear directional mean and cosine similarity*

Though Model 2 has the advantage of simplicity, it cannot handle two types of linear feature matched pairs when two lines in each pair are in large angle difference, usually intersecting at an angle close to 90°: (1) one line is considerably shorter than the other (Figure 3; Table 2); and (2) the median distance between the two lines is shorter than a threshold (Figure 4). Indeed, direction and angle have been identified in Tong *et al.* (2014) to account for mismatched cases. In their study area, these incorrectly matched pairs appear to be a minor issue, but the mismatched cases can be more pronounced in studies involving historical data sets and curved road networks, such as this study.

To address the deficiency in Model 2, Model 3 compares the angles of the matched pairs for the intersected parts within a buffer polygon because the entire lines can be in varying angles but the intersected parts are likely to be in a similar angle. For example,

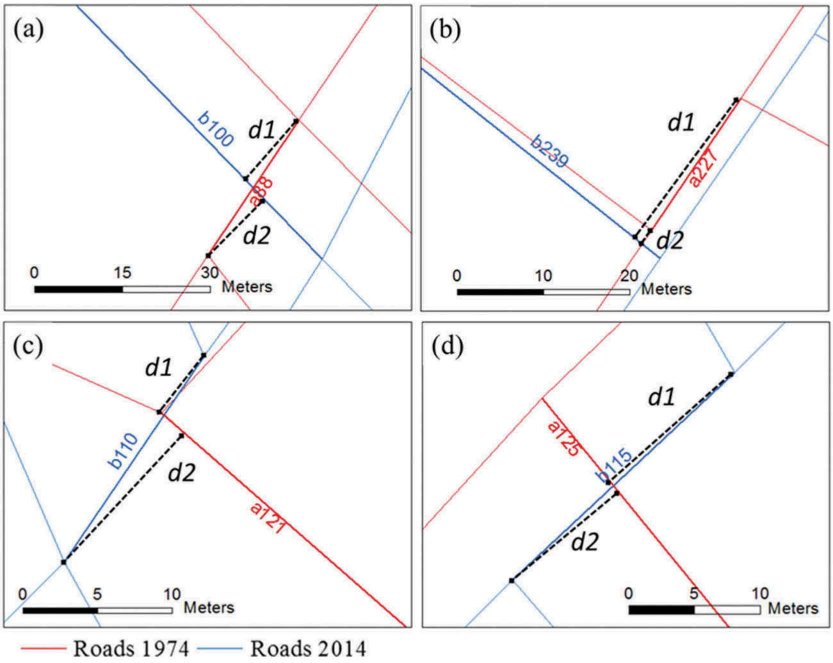

Figure 3. Incorrectly matched pairs in case that one line is considerably shorter than the other line. The value differences are provided in Table 2.

Table 2. The values of the length difference, angle, d1, d2, and SMHD in Figure 3.

	Shorter		Longer						
Matched pair	ID	Length	ID	Length	ΔLength	ΔAngle	d1	d2	SMHD
a	a88	27.74	b100	125.29	97.55	78	13.97	13.67	13.82
b	a227	18.55	b239	212.50	193.95	86	20.80	2.04	11.42
c	b110	16.65	a121	28.22	11.57	83	11.73	4.70	8.20
d	b115	23.12	a125	104.71	81.59	86	12.80	10.25	11.53

The units for all variables are in meter, except for angle, which is degree. Matched pairs a–d, respectively, correspond to Figure 3(a–d).

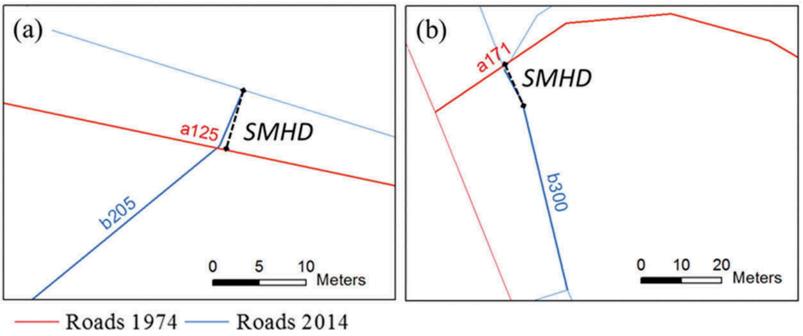

Figure 4. Incorrectly matched pairs where the median distance between the two lines is shorter than a threshold (i.e., 18.31 m in this case). SMHD is 6.68 m in (a) and 9.97 m in (b).

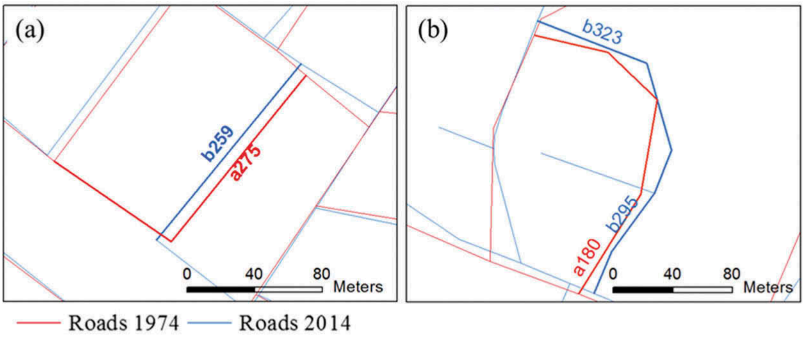

Figure 5. Partially matched pairs.

b259 in Figure 5(a) is partially matched to a275, and a180 in Figure 5(b) is matched to both b295 and b323. In case of Figure 5(b), b295 is in different angles with respect to the entire a180. The partial matching between b295 and part of a180 intersected with the b295 buffer polygon will therefore be compared for angles.

Before the comparison of the angles of the linear features, the linear directional mean (LDM) of each linear feature is calculated because the directions of line segments are not invariant (Equation 4).

$$LDM = atan\frac{\frac{1}{n}\sum_{i=1}^{n}\sin(\theta_i)}{\frac{1}{n}\sum_{i=1}^{n}\cos(\theta_i)} = atan\frac{\sum_{i=1}^{n}\sin(\theta_i)}{\sum_{i=1}^{n}\cos(\theta_i)}, \tag{4}$$

where θ_i is the direction of line segment i, and n is the number of line segments.

Given the quadrant that the sum of sine and cosine values belong to, the LDM in Equation (4) is adjusted to represent the degree measured counterclockwise from 0° (Equation 5) (Wong and Lee 2005).

$$\begin{aligned}
&\sum_{i=1}^{n}\sin(\theta_i) \geq 0 \; and \sum_{i=1}^{n}\cos(\theta_i) \geq 0 \; LDM = LDM \\
&\sum_{i=1}^{n}\sin(\theta_i) \geq 0 \; and \sum_{i=1}^{n}\cos(\theta_i) < 0 \; LDM = 180 - |LDM| \\
&\sum_{i=1}^{n}\sin(\theta_i) < 0 \; and \sum_{i=1}^{n}\cos(\theta_i) < 0 \; LDM = 180 + LDM \\
&\sum_{i=1}^{n}\sin(\theta_i) < 0 \; and \sum_{i=1}^{n}\cos(\theta_i) \geq 0 \; LDM = 360 - |LDM|.
\end{aligned} \tag{5}$$

For more accurate calculation of the direction difference, weights for line segment lengths are applied to the LDM. Thus, Equation (4) is modified as Equation (6), which is the WLDM:

$$WLDM = atan\frac{\sum_{i=1}^{n}\sin(\theta_i) \times \frac{l_i}{l_t}}{\sum \cdot_{i=1}^{n}\cos(\theta_i) \times \frac{l_i}{l_t}}, \tag{6}$$

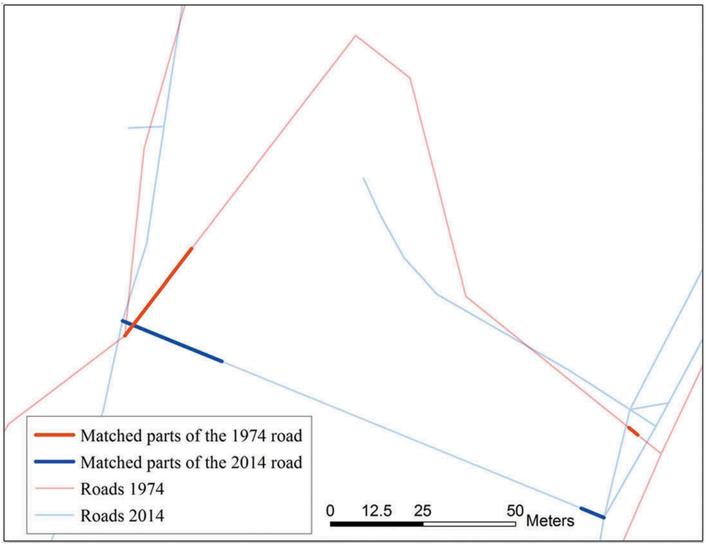
Matched parts of the 1974 road
Matched parts of the 2014 road
Roads 1974
Roads 2014
0 12.5 25 50 Meters

Figure 6. Partially matched pairs with more than one partial overlap segment. The longer line segments, highlighted using thicker lines in the left-hand side of the figure, are in different directions. The shorter line segments, highlighted using thicker lines in the bottom-right corner, are in the similar directions.

where θ_i is the direction of line segment i, l_i is the length of the line segment i, and l_t is the total length. For example, in Figure 6, the longer line segments of 1974 and 2014 linear features are in different directions, but the shorter line segments are in similar directions. The WLDM can show that these two sets of line segments are in different directions, and provide more accurate direction description for partially overlapping line segments than the direction comparison based on a start point and an end point. In the case of Figure 6, the difference of directions between the start point and the end point is 11.5°, but the direction difference using the WLDM is 68.6°.

Once the direction difference (θ) is calculated by comparing the WLDMs, it is converted to a cosine similarity (CS) value. The CS is a similarity measure based on the cosine value of the direction difference between two vectors:

$$\mathrm{CS} = \cos(\theta). \tag{7}$$

The CS ranges between −1 and 1. Values close to 1 and −1, respectively, indicate that two vectors are in similar and opposite directions. The value of 0 denotes two perpendicular vectors. Because of the lack of the direction attribute in many of the historical road data, we calculate angle differences between two lines. Consequently, the aCS is used in this study, with the values ranging from 0 to 1.

Based on the matching results from Model 2, a second decision tree is constructed to derive the threshold for the aCS. The aCS can be used to check the many-to-many matching candidates collected from Model 2. If the aCS of the candidates is above the threshold, it is highly likely that the matching candidates are correct.

Model 3 is based on both the SMHD and the aCS thresholds, but it may treat certain correctly matched pairs as incorrectly matched pairs because of the thresholds. To tackle this issue, Model 3 is divided into two sub-models, Model 3–1 (Equation 8) and Model 3–2 (Equation 9) (Figure 1). The input of Model 3–1 is the matched pairs generated by Model 2, while the input of Model 3–2 is the combined outputs of the final matched pairs from Model 1 and Model 2. Model 3–1 follows the SMHD threshold with the matched pairs generated by Model 2. Conversely, Model 3–2 integrates the matched pairs in Model 1 that exceed the SMHD threshold. In Model 1, the one-to-one matched pairs are classified by their closeness to the counterpart lines, so some of them can potentially be correctly matched pairs. Model 3–2 is thus designed to detect such pairs.

$$\mathrm{Model3}-1=\begin{cases} T, \mathit{if\ SMHD} \leq \mathit{a\ threshold\ and\ aCS} \geq \mathit{a\ threshold} \\ F, \mathit{otherwise} \end{cases}. \tag{8}$$

$$\mathrm{Model3}-2=\begin{cases} T, \mathit{if\ SMHD} \leq \mathit{a\ threshold\ and\ aCS} \geq \mathit{a\ threshold} \\ T, (\mathit{if\ Matching} = T\ \mathit{in\ Model\ 1\ or\ SMHD} \leq \mathit{a\ threshold}) \\ \quad \mathit{and\ aCS} \geq \mathit{a\ threshold} \\ F, \mathit{otherwise} \end{cases}. \tag{9}$$

2.4. *Model 4: topological relationships*

Although the aCS is useful to identify a similar linear shape, some pairs can be incorrectly matched because they are in opposite directions from the matched junctions. To solve this issue, topological relationships are considered. To investigate topological relationships, matched junctions are identified with the road junction data set created in data preprocessing. If two road junctions from the target data set and the reference data set share more than three matched linear features connected to each junction, they are considered as a matched junction pair. If the start point of a linear feature is not linked to the matched junction, the line is flipped so that the directions from the matched junctions can be calculated. The incorrectly matched pairs with opposite directions from the matched junctions but similar aCS values are also of two types: (1) one line is considerably shorter than the other line (Figure 7); and (2) the median distance between the two lines is shorter than a threshold (Figure 8).

In Figure 7(a), based on Model 3, a161 is matched to b150. The SMHD is the mean value of two distances from the end points of the shorter line to the counterpart line. Although the pair satisfies both the SMHD and the aCS thresholds, it is incorrect due to the opposite direction from the matched junctions (P_a and P_b). In the reference data set, a matched linear feature of a161 does not exist because two three-way junctions of the target data set are merged to a four-way junction. The thresholds of the SMHD and the aCS cannot identify such error. Figure 7(b) illustrates a similar case that a163 and b145 are an incorrectly matched pair because a163 and b145 are in opposite directions from the matched junctions (P_a and P_b).

The incorrectly matched pairs in Figure 8 are related to the median distance shorter than a threshold. In Figure 8(a), b108 appears to be a straight line with two end points, but there is one vertex on the line. Due to the vertex, the SMHD is the distance from the vertex to a107, and the pair is classified as a matched pair by Model 3. However, two lines are in opposite directions from the matched junctions (P_a and P_b). In Figure 8(b), b204 is the shorter line comprised of four vertices. In this case, the SMHD is the mean of

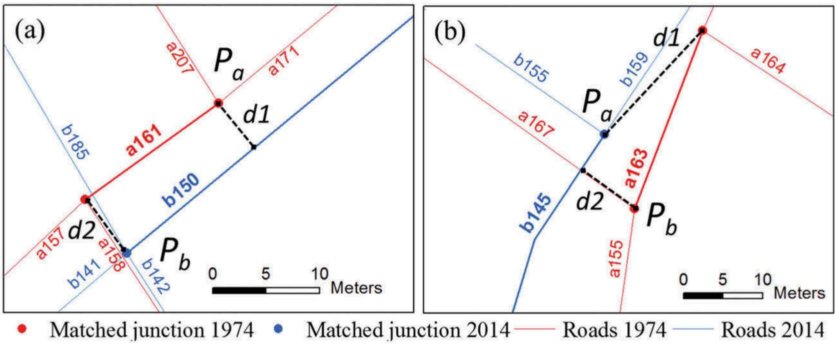

Figure 7. Incorrectly matched pairs from the matched junctions where one line is considerably shorter than the other. SMHD = $(d_1 + d_2)/2$.

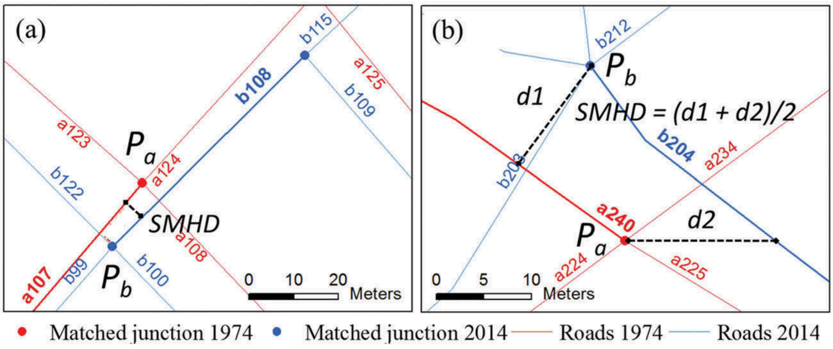

Figure 8. Incorrectly matched pairs from the matched junctions where the median distance between the two lines is shorter than a threshold.

two middle distance values and shorter than a threshold. The aCS of the matched pair is above the aCS threshold, but the matched features are also in opposite directions from matched junctions (P_a and P_b).

The solution to such pairs is to compare topological relationships on the matched junctions with three steps: (1) search for neighboring linear features; (2) investigate whether the junctions between the linear features and the neighboring linear features are matched junctions; and (3) examine whether the linear features are in a similar direction from the matched junctions. In Figure 9, the matched pair p_1 and q_1 is selected, and whether there is another matched pair of q_1 is searched. If another pair exists (p_2) (Figure 9(b)), it is investigated whether p_1 and p_2 are neighboring features and whether a junction exists between the two linear features p_1 and p_2. The same process is then applied to q_1 to identify if it has a neighboring feature. If there is no neighboring feature to q_1, the end point of q_1 that is the closest to the junction between p_1 and p_2 is considered as a matched junction candidate (Figure 9(b)). Alternatively, if there is a neighboring feature to q_1, then two junctions are identified (Figure 9(c)), and more lines are collected from the junction tables. If more lines connected to the junction are found, additional linear feature matched pairs are searched (p_3, p_4, q_3, and q_4) (Figure 9(d)). If more than three matched pairs including the original matched pair exist, two junctions are considered as a matched junction pair.

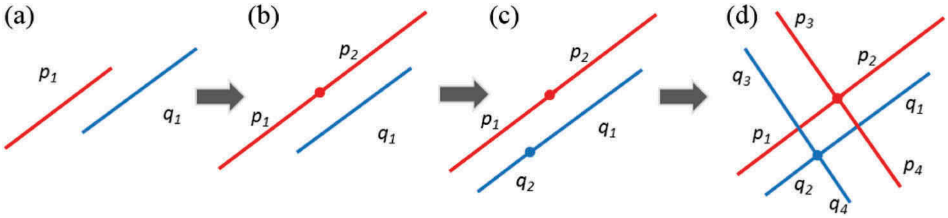

Figure 9. The procedure to identify matched junctions of a matched pair.

Once the junctions are confirmed as a pair, the directions of the matched linear features and neighboring features are adjusted so that their vertices start from the junctions before the comparison of the directions. In Figure 10, α, β, and γ are the directions of p_1, p_2, and q_1, respectively, with p_1, and q_1 in opposite directions from the junction and |α – γ |greater than |β – γ | (Equation 10). Model 4 is divided into Model 4–1 and Model 4–2, each of which has its corresponding input from Model 3 (Figure 1). The input of Model 4–1 is the matched pairs of Model 3–1, and that of Model 4–2 is the matched pairs of Model 3–2. The topological relationships are applied to both Model 4–1 and Model 4–2.

$$Topological\ matching = \left\{ \begin{matrix} T, |\alpha - \gamma| \leq |\beta - \gamma| \\ F, |\alpha - \gamma| > |\beta - \gamma| \end{matrix} \right\}. \tag{10}$$

2.5. *Model evaluation*

To evaluate the model performance, all pairs and non-pairs are first examined manually with the aerial photos in the 1950s and the current satellite image from the ArcGIS® World Imagery Basemap. Then, accuracy (Equation 11) and recall (Equation 12) are calculated to compare the matching outcome from each model with the pairs:

$$\text{Accuracy} = \frac{Correct\ Pairs}{Correct\ Pairs + Incorrect\ Pairs} \times 100\% \tag{11}$$

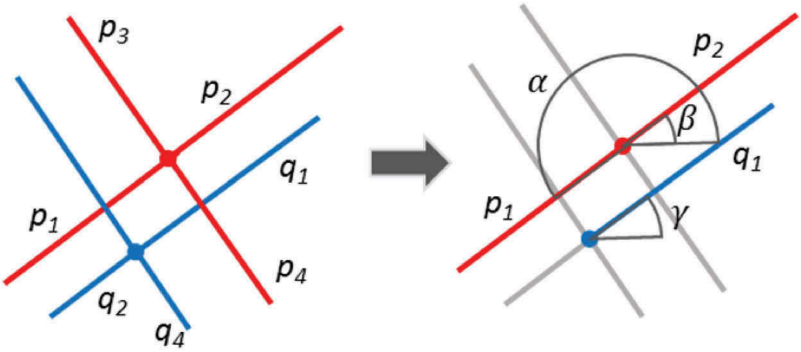

Figure 10. The directions of matched features from matched junctions.

$$\text{Recall} = \frac{\textit{Correct Pairs}}{\textit{Correct Pairs} + \textit{Incorrect Nonpairs}} \times 100\%, \qquad (12)$$

where the terms refer to the number of linear features correctly or incorrectly classified as pairs or non-pairs by each model.

3. Experiments and discussion

The experiments with all four models are conducted in the Singapore downtown with two test areas (Figure 11). The two test areas are selected for their differences in linear feature characteristics. Test Area 1 contains predominantly straight linear features, while Test Area 2 contains more curved features (Figure 12). The models are carried out in the two test areas before being applied to the whole Singapore downtown, so as to: (1) build and test the models with less computing time; (2) examine the performance differences based on the complexity (i.e., straight linear features vs. curved linear features); and (3) examine the influence of spatial variation of thresholds on performance. For the Singapore downtown, two matching experiments are performed: with and without multilane expressways. Table 3 summarizes the road network data sets for the test areas and the Singapore downtown.

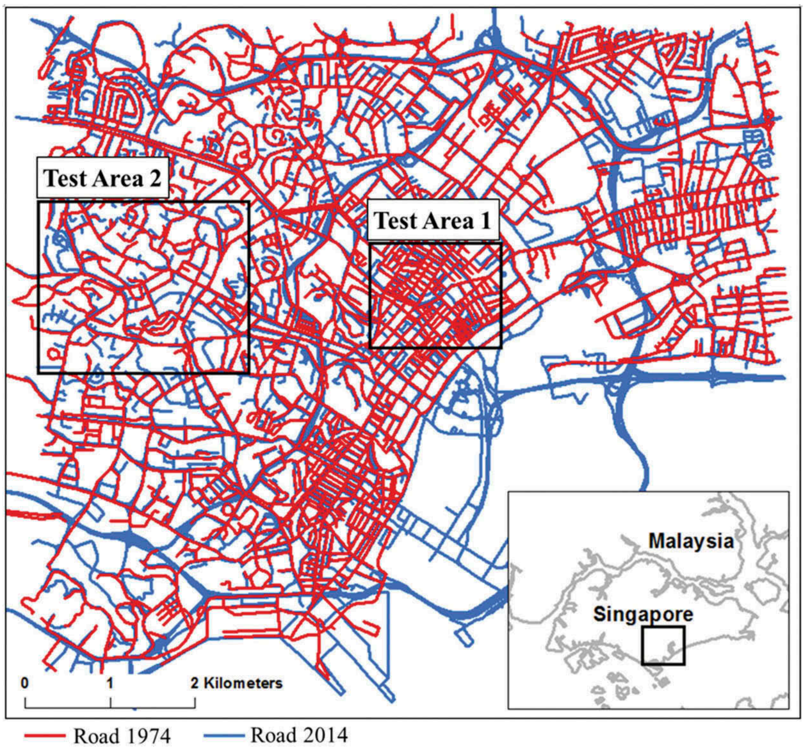

Figure 11. The road network data of the Singapore downtown.

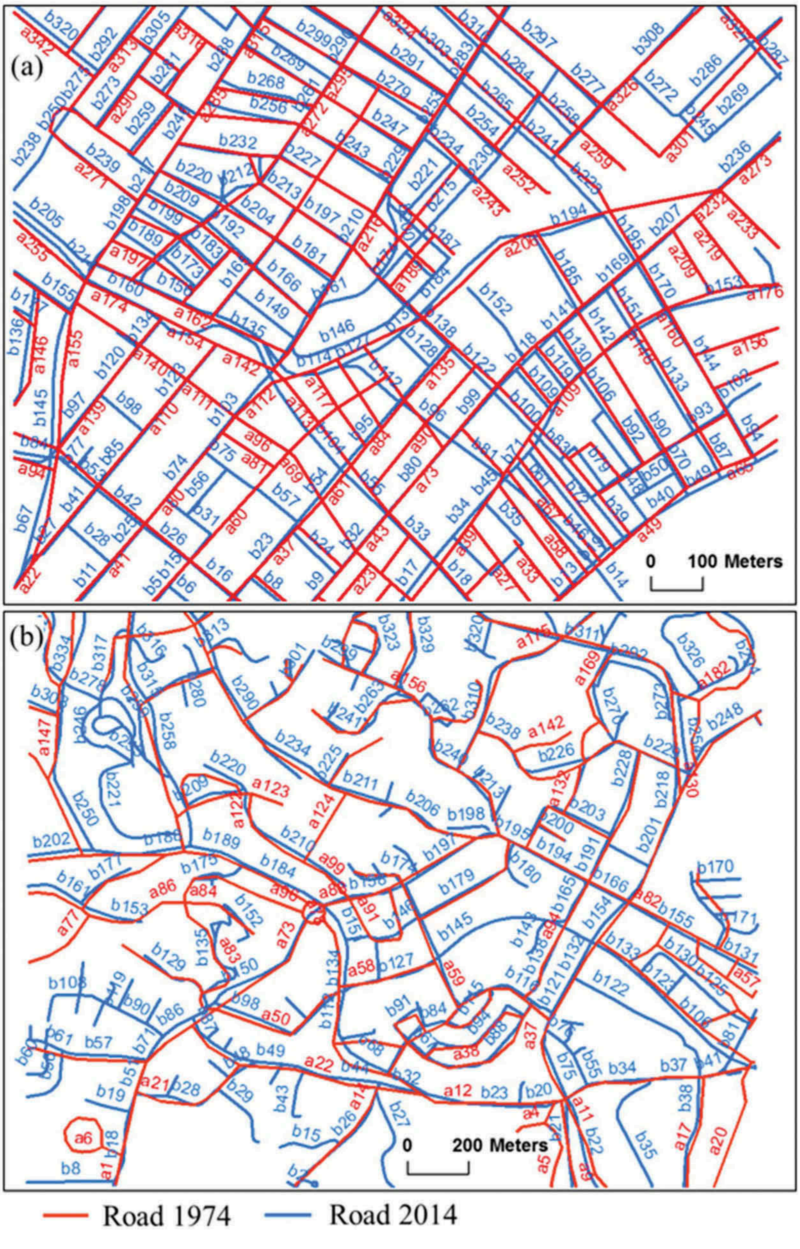

Figure 12. The road network data of (a) Test Area 1 and (b) Test Area 2.

3.1. *Matching results and evaluation*

For the thresholds of the SMHD, 16.19 m, 18.31 m, and 14.95 m are yielded for Test Area 1, Test Area 2, and the Singapore downtown, respectively. The thresholds of the aCS are 0.9435 (19.34°), 0.9735 (13.22°), and 0.9558 (17.1°) for Test Area 1, Test Area 2, and the Singapore downtown, respectively. The matching results and the performances of all models for Test Area 1, Test Area 2, and the Singapore downtown are illustrated in Tables 4–7.

Table 3. Summary of the target and reference data sets for the study areas.

	Test Area 1		Test Area 2		Singapore downtown	
	Target	Reference	Target	Reference	Target	Reference
Number of lines	334	315	190	327	2927	6090 (5475)
Total length (km)	31	29	36	44	453	686 (588)

The values in the brackets for the Singapore downtown reference data set denote the number of lines and lengths excluding multilane features.

Both Model 1 and Model 2 consider distance. Model 1 is one-to-one matching based on the SMHD only, while Model 2 is many-to-many matching using a threshold for the SMHD. Comparison of model performances reveals that Model 2 has higher accuracy at 82.3% than Model 1 at 79.4% for Test Area 1. But for other areas, Model 1 outperforms Model 2 (i.e., the accuracy for Model 1 vs. Model 2 is 79.7% vs. 72.2% in Table 5, 71.9% vs. 65.0% in Table 6, and 75.1% vs. 70.5% in Table 7). These results suggest that considering distance alone can give rise to unstable performance. Unlike the accuracy, the recall of Model 2 is higher than that of Model 1 for all study areas (Tables 4–7) because the linear features classified as incorrect non-pairs by Model 1 have been included.

The incorporation of distance (i.e., the SMHD) and angle (i.e., the aCS) into Model 3 produces better and more stable performances than Model 2, evident in the accuracies at around 90% for Model 3–1 and Model 3–2 for all study areas. For instance, for Model 3–1, the accuracy for Test Area 2 shows an increase of 19.2% from 72.2% to 91.4% (Table 5), and for Singapore downtown area, an increase of 24.4% from 65.0% to 89.4% (Table 6). These results indicate that angle is an important component to stabilize the performance of the matching method, and that higher accuracy can be achieved by incorporating distance and angle.

Because Model 3 can wrongly classify the linear features in opposite directions from a matched junction pair as correct pairs, to further improve the accuracy, junction matching is incorporated into Model 4 to consider the topological relationships, as opposed to simply

Table 4. Performance of Models 1–4, Test Area 1.

	Model 1	Model 2	Model 3–1	Model 3–2	Model 4–1	Model 4–2
Total pairs	321	339	317	338	291	309
Correct pairs	255	279	279	295	278	294
Incorrect pairs	66	60	38	43	13	15
Incorrect non-pairs	46	22	22	6	23	7
Accuracy (%)	79.4	82.3	88.0	87.3	95.5	95.1
Recall (%)	84.7	92.7	92.7	98.0	92.4	97.7

Table 5. Performance of Models 1–4, Test Area 2.

	Model 1	Model 2	Model 3–1	Model 3–2	Model 4–1	Model 4–2
Total pairs	187	263	197	209	182	192
Correct pairs	149	190	180	189	177	186
Incorrect pairs	38	73	17	20	5	8
Incorrect non-pairs	53	12	19	10	22	13
Accuracy (%)	79.7	72.2	91.4	90.0	97.3	95.9
Recall (%)	73.8	94.1	90.5	95.0	89.0	93.5

Table 6. Performance of Models 1–4, the Singapore downtown.

	Model 1	Model 2	Model 3–1	Model 3–2	Model 4–1	Model 4–2
Total pairs	2837	4012	2909	3283	2723	3091
Correct pairs	2041	2606	2601	2920	2585	2904
Incorrect pairs	796	1406	308	363	138	187
Incorrect non-pairs	1283	718	723	404	739	408
Accuracy (%)	71.9	65.0	89.4	88.9	95.0	94.0
Recall (%)	61.4	78.4	78.2	87.8	77.8	87.4

Table 7. Performance of Models 1–4, the Singapore downtown when multilane features were excluded.

	Model 1	Model 2	Model 3–1	Model 3–2	Model 4–1	Model 4–2
Total pairs	2584	3375	2610	2967	2655	2789
Correct pairs	1940	2380	2376	2689	2585	2681
Incorrect pairs	644	995	234	278	70	108
Incorrect non-pairs	1071	631	635	322	739	330
Accuracy (%)	75.1	70.5	91.0	90.6	97.4	96.1
Recall (%)	64.4	79.0	78.8	89.3	77.8	89.0

employing shape similarity (i.e., the SMHD and the WLDM). Consequently, the performances of Model 4–1 and Model 4–2 improve, with most of the study areas achieving at least 95% of the accuracy. For example, the accuracy for Test Area 1 increases from 87.3% for Model 3–2 to 95.1% for Model 4–2 (Table 4). The accuracy for Test Area 2 improves from 91.4% for Model 3–1 to 97.3% for Model 4–1 (Table 5). The recall of Model 4–2 is higher than that of Model 4–1 because Model 4–2 includes more potential matched pairs beyond the SMHD threshold. For the Singapore downtown area, slight increase in accuracy and recall are observed when multilane features of expressways are excluded from the calculation. Nevertheless, as there are only 2–3% differences in accuracies between the results with and without the multilane features, the similar model performances for the large data set of the Singapore downtown area (Tables 6–7) confirm the robustness of the method.

The experiments demonstrate that our proposed method improves upon previous approaches in at least three aspects. First, our method automatically generates the optimal thresholds. This is different from Zhang and Meng (2007) and Yang *et al.* (2014) in that their thresholds are defined by researchers, which might not be the optimal solutions, and potentially weakens the model performance because they can be parameter sensitive. In contrast to the methods of Yang *et al.* (2013) and Tong *et al.* (2014) that employ various iterative methods to automate the threshold calculations, our method does not rely on iterations to derive the optimal thresholds. Second, our method employs a simple rule to compare topological relationships, rather than depending on expert knowledge to classify complicated road junctions. This has the advantage over prior work, such as Zhang and Meng (2007), in which researchers have to identify the diverse road junction types based on their prior knowledge so as to define a variety of topological rules. The topological rule of our method simply compares the number of matched linear features. Third, even though our method is simpler than previous approaches, our method can process curved linear features and still guarantees robust results. This is supported by the improved accuracies of Model 3 (88.0% - 91.4%) which considers distance and direction, and of Model 4 (94.0% - 97.4%) which further incorporates topological relationships.

3.2. *Incorrectly matched pairs and possible solutions*

Despite the improved and robust performances of our method, there are still a few incorrectly matched pairs related to the threshold-based matching method. One issue is caused by the different LoDs of two data sets. The expressways are represented as single lanes in the target data set and multilanes in the reference data set. Some of the multilanes are beyond the SMHD thresholds from the corresponding single lane, so they have been incorrectly matched. The other issue is due to changes in road network configurations, particularly the typology of junctions (Figure 13(a,b)). In Figure 13(a), a708 is split into b959 and b1048 by a new junction, but b959 is not matched to a708 because the SMHD of the pair is greater than the threshold. In Figure 13(b), the T junction in the target data set has changed to the Y junction in the reference data set. Intuitively, b3225 can be either matched to both a2256 and a2228, or b3225 finds no matching linear feature in the target data set. Yet, b3225 is matched to a2256 while a2228 is not matched to b3225 because the SMHD between two linear features is greater than the SMHD threshold.

A solution to these issues is to increase the threshold so as to include the potential correct pairs that would have otherwise been classified into incorrect non-pairs. Nonetheless, in so doing, additional incorrect pairs will be added, possibly decreasing the model performance. Indeed, when we adopt a larger threshold than the original threshold derived from the decision tree analysis, the model performance degrades, with the accuracy of Model 3 down to approximately 80%. As Yang *et al*. (2014) have noted, the threshold-based method cannot classify all of the

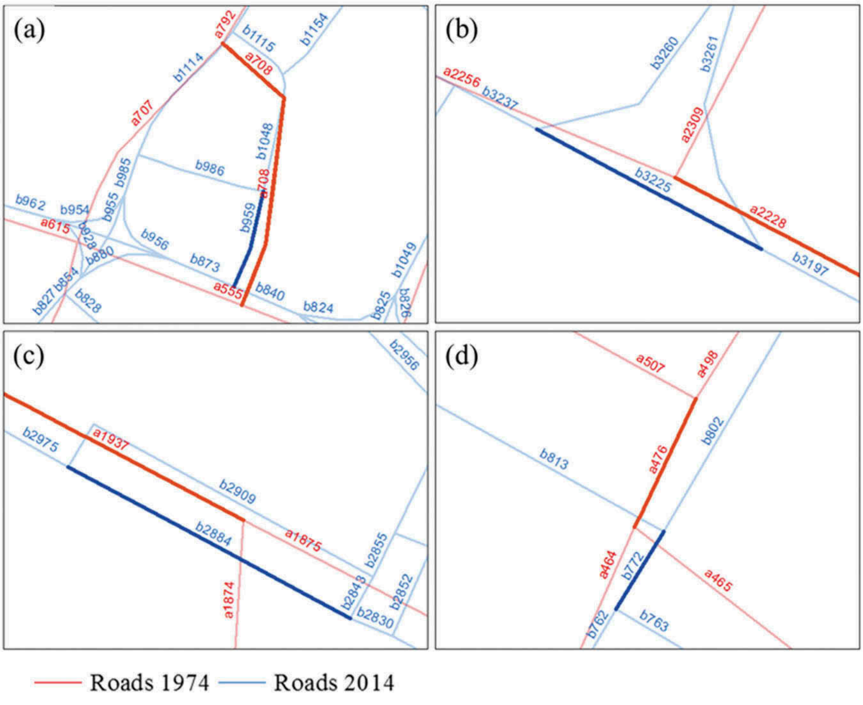

Figure 13. The examples of incorrectly matched pairs.

matched pairs perfectly. Given that the decision tree analysis has provided the optimal threshold, further accuracy improvement will require variables independent from thresholds. For example, additional topological rules can be defined to examine features adjacent to matched pairs. In the case of Figure 13(a), b959 can be reexamined with other new criteria because it is connected to a matched pair (i.e., b1048). Due to the various types of incorrectly matched cases, it would, however, be challenging to apply the same topological rules to match all of them (Zhang and Meng 2007).

Positional uncertainty (Figure 13(c,d)) is another reason accounting for the incorrectly matched pairs. The lack of projection information and digitizing errors in the target data set lead to the positional discrepancies with the reference data set. The positional error tolerance of our study is 25 m, so all roads from the reference data set within the 25 m buffer of the target data set are considered as the matching candidates. The correct counterpart road in the reference data set may hence not be the closest to the road in the target data set. It is also possible that a road of the target data set is matched to two roads of the reference data set located on its two sides, particularly when these two roads bound a narrow, elongated block. For example, in Figure 13(c), a1937 is closer to b2909, but the correct counterpart road is b2884. In Figure 13(d) the topological relationship in Model 4 is investigated based on the closest junction for the matched pairs. Thus, the junction between a476 and a464 is compared with the junction between b802 and b772, instead of comparing with the junction between b762 and b772, which is the correctly matched junction.

To minimize positional errors, the use of ancillary data (Wong *et al*. 2012) and evolutionary genetic algorithms to generate a Pareto front of solutions (Manzano-Agugliaro *et al*. 2013) have been proposed. Most of these methods reduce the gross error, while approaches to minimize the local errors will be desirable. Besides, obtaining ancillary data and redefining the spatial references are often more challenging for historical maps than contemporary data (Tucci and Giordano 2011). One practical solution is to incorporate positional error assessment methods with the tools visualizing the spatial variation of positional errors, such as interpolated displacement vectors (Seo and O'Hara 2009) and statistical simulation models for positional errors (Tong *et al*. 2013). By providing user-interfaces to visualize positional errors, users will be able to inspect the areas of large errors and find incorrectly matched pairs easily.

4. Conclusion

This study presents a simple yet robust linear feature matching method to conflate historical and current road data, based on shape similarity and topological relationships. Four models with incremental configurations are tested to demonstrate the usefulness of the proposed method. The use of decision tree analysis has the advantage to automatically obtain the optimal thresholds for the classification of correctly and incorrectly matched pairs. The results show that considering distance alone cannot warrant stable model performances, while our proposed Model 3, which considers shape similarity, provides stable matching results, achieving around 90% accuracy for all study areas. Model 4 further improves the accuracy to 95% with a simple topological rule. The proposed method is not only simpler than previous approaches but also can process curved linear road features to give robust performances. Just as all other threshold-base matching approaches, the proposed method is not without limitation. Efforts on reducing the positional errors of the historical data will be desirable to improve the matching accuracy.

Acknowledgments

We would like to thank the anonymous reviewers and the editor for their insightful comments on the paper, as we could improve our paper based on the comments.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This work was supported by the Ministry of Education - Singapore Tier 2 Grants [MOE2015-T2-2-135 and WBS: R-102-000-101-112] and the Strategic Initiative of the National University of Singapore [Grant Number: WBS: C-109-000-025-001].

ORCID

Ick-Hoi Kim http://orcid.org/0000-0003-4034-8530
Chen-Chieh Feng http://orcid.org/0000-0003-0410-714X
Yi-Chen Wang http://orcid.org/0000-0002-3034-7377

References

Beeri, C., *et al*., 2004. Object fusion in geographic information systems. *In*: M.A. Nascimento, *et al*., eds. *Proceedings of the thirtieth international conference on very large data bases*, 31 August–3 September 2004 Toronto, Canada. San Fransisco, CA: Morgan Kaufmann, 816–827.

Du, H., *et al*., 2012. Geospatial information integration for authoritative and crowd sourced road vector data. *Transactions in GIS*, 16 (4), 455–476. doi:10.1111/j.1467-9671.2012.01303.x

Hastings, J.T., 2008. Automated conflation of digital gazetteer data. *International Journal of Geographical Information Science*, 22 (10), 1109–1127. doi:10.1080/13658810701851453

Hope, S. and Kealy, A., 2008. Using topological relationships to inform a data integration process. *Transactions in GIS*, 12 (2), 267–283. doi:10.1111/j.1467-9671.2008.01098.x

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. doi:10.1080/19479832.2011.577458

Manzano-Agugliaro, F., *et al*., 2013. Pareto-based evolutionary algorithms for the calculation of transformation parameters and accuracy assessment of historical maps. *Computers & Geosciences*, 57, 124–132. doi:10.1016/j.cageo.2013.04.010

Quinlan, J.R., 1986. Induction of decision trees. *Machine Learning*, 1, 81–106. doi:10.1007/BF00116251

Quinlan, J.R., 1993. *C4.5: Programs for Machine Learning*. San Mateo, CA: Morgan Kaufmann. doi:10.1016/S0019-9958(62)90649-6

Saalfeld, A., 1988. Conflation automated map compilation. *International Journal of Geographical Information Systems*, 2 (3), 217–228. doi:10.1080/02693798808927897

Safra, E., *et al*., 2010. Location-based algorithms for finding sets of corresponding objects over several geo-spatial data sets. *International Journal of Geographical Information Science*, 24 (1), 69–106. doi:10.1080/13658810802275560

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. doi:10.1080/13658810410001658076

Seo, S. and O'Hara, C.G., 2009. Quality assessment of linear data. *International Journal of Geographical Information Science*, 23 (12), 1503–1525. doi:10.1080/13658810802231456

Shannon, C.E., 1948. A mathematical theory of communication. *The Bell System Technical Journal*, 27, 379–423. doi:10.1002/j.1538-7305.1948.tb01338.x

Tong, X., *et al*., 2013. A statistical simulation model for positional error of line features in Geographic Information Systems (GIS). *International Journal of Applied Earth Observation and Geoinformation*, 21 (1), 136–148. doi:10.1016/j.jag.2012.08.004

Tong, X., Liang, D., and Jin, Y., 2014. A linear road object matching method for conflation based on optimization and logistic regression. *International Journal of Geographical Information Science*, 28 (4), 824–846. doi:10.1080/13658816.2013.876501

Trainor, T., 2003. U.S. Census Bureau geographic support: a response to changing technology and improved data. *Cartography and Geographic Information Science*, 30 (2), 217–223. doi:10.1559/152304003100011054

Tucci, M. and Giordano, A., 2011. Positional accuracy, positional uncertainty, and feature change detection in historical maps: results of an experiment. *Computers, Environment and Urban Systems*, 35 (6), 452–463. doi:10.1016/j.compenvurbsys.2011.05.004

Van Niel, T.G. and McVicar, T.R., 2002. Experimental evaluation of positional accuracy estimates from a linear network using point- and line-based testing methods. *International Journal of Geographical Information Science*, 16 (5), 455–473. doi:10.1080/13658810210137022

Walter, V. and Fritsch, D., 1999. Matching spatial data sets: a statistical approach. *International Journal of Geographical Information Science*, 13 (5), 445–473. doi:10.1080/136588199241157

Wong, D.W., So, B.K.L., and Zhang, P., 2012. Addressing quality issues of historical GIS data: an example of Republican Beijing. *Annals of GIS*, 18 (1), 17–29. doi:10.1080/19475683.2011.647074

Wong, D.W.S. and Lee, J., 2005. *Statistical analysis of geographic information with ArcView GIS and ArcGIS*. Hoboken, NJ: John Wiley & Sons, Inc.

Yang, B., Luan, X., and Zhang, Y., 2014. A pattern-based approach for matching nodes in heterogeneous urban road networks. *Transactions in GIS*, 18 (5), 718–739. doi:10.1111/tgis.12057

Yang, B., Zhang, Y., and Luan, X., 2013. A probabilistic relaxation approach for matching road networks. *International Journal of Geographical Information Science*, 27 (2), 319–338. doi:10.1080/13658816.2012.683486

Yuan, S. and Tao, C., 1999. Development of conflation components. *In*: *The proceedings of geoinformatics'99 conference*, 19–21 June 1999 Ann Arbor, MI. 1–13. Available from: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.116.3218&rep=rep1&type=pdf.

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. doi:10.1016/j.compenvurbsys.2007.08.008