DOI: 10.1111/tgis.12726

RESEARCH ARTICLE

Transactions in GIS WILEY

Conflating linear features using turning function distance: A new orientation-sensitive similarity measure

Ting L. Lei[1] | Rongrong Wang[2,3]

[1]Department of Geography and Atmospheric Science, University of Kansas, Lawrence, KS, USA

[2]Department of Computational Mathematics, Science and Engineering, Michigan State University, East Lansing, MI, USA

[3]Department of Mathematics, Michigan State University, East Lansing, MI, USA

Correspondence
Rongrong Wang, Department of Computational Mathematics, Science and Engineering, Michigan State University, East Lansing, MI 48824, USA.
wangron6@msu.edu

Funding information
National Natural Science Foundation of China, Grant/Award Number: 41971334; NSF CCF, Grant/Award Number: 1909523

Abstract

Measuring the similarity between counterpart geospatial features is crucial in the effective conflation of spatial datasets from difference sources. This article proposes a new similarity metric called the "map turning function distance" (MTFD) for matching linear features such as roads based on the well-known turning function (TF) distance in computer vision. The MTFD overcomes the limitations of the traditional TF distance, such as the inability to handle partial matches and insensitivity to differences in scale and rotation. In particular, the MTFD allows one to: (a) partially match a linear feature to a portion of a larger feature from a certain position of match; and (b) consider both the shape and orientation differences of polylines based on comparing their turning angles. In finding the best match position, we prove that the optimal position can be found among a finite set of positions on the target feature. We then combine the MTFD with widely used point-offset distances such as the Hausdorff distance to form a composite similarity metric. Our experiments with real road datasets demonstrate that the new metric has greater discriminative power than traditional point-offset-based similarity measures, and significantly improves the precision of two tested conflation models.

1 | INTRODUCTION

Conflation is needed in a wide range of spatial analyses because data about a spatial phenomenon are often produced by different agencies and vendors, each with a different role and scope. Take road networks for example. Public agencies such as the U.S. Census and the U.S. Geological Survey maintain regularly updated databases such as TIGER/Line (with rich socioeconomic attributes). Private companies such as TomTom and HERE provide network datasets for navigation and planning purposes, with better geometric accuracy and attributes of transportation infrastructure. Open data projects, such as OpenStreetMap (OSM), have made an increasingly richer set of data available as volunteered geographic information (VGI). Transportation research often needs comprehensive information from all these sources in order to study travel behavior, predict traffic flow, and develop future policies and plans. For this reason, conflation has an important role in transportation studies.

In conflating linear features such as road segments, a fundamental question is how to effectively measure the similarity between features. Traditionally, offset-based metrics, such as the Hausdorff distance, have been used widely to measure the difference between coordinates of a pair of features (e.g., Chehreghan & Abbaspour, 2017b; Li & Goodchild, 2010, 2011; Tong, Liang, & Jin, 2014; Xavier, Ariza-López, & Ureña-Cámara, 2016). However, Hausdorff-like distances can be unstable, especially in matching small features. As will be discussed in the next section, due to its size, a small linear feature can often be matched to any nearby polylines without regard to its shape and orientation, resulting in erroneous matches.

In this article, we propose to measure the (dis)similarity between two linear features by considering their angular differences in terms of shape (deformation) and orientation (rotation), respectively. Inspired by the standard turning function (TF) distance (Arkin, Chew, Huttenlocher, Kedem, & Mitchell, 1991) between polygons, we develop an orientation-sensitive "map turning function distance" (MTFD) for measuring the discrepancies of linear features in terms of angular differences. We prove that the correct shape deformation and rotation can be found by evaluating the difference of directional angles at a finite number of match positions on the target feature (called *critical event positions*). We then combine the new angular similarity measure with the conventional offset-based Hausdorff distance. We evaluate the effectiveness of the combined similarity measure with case studies and performance tests using two open datasets in Santa Barbara, CA from OSM and TIGER/Line, respectively.

2 | BACKGROUND

In this section, we briefly review related work and concepts for linear feature conflation, with a focus on geometric similarity measures and typical match-selection methods. The interested reader is referred to Ruiz, Ariza, Ureña, and Blázquez (2011) and Xavier et al. (2016) for comprehensive reviews on other aspects of the conflation problem.

2.1 | Measures of similarity between features

A fundamental concept for matching geographic features is the closeness or similarity between a pair of features. It indicates the likelihood that the two features may belong to the same object in reality. Effective conflation relies on good similarity measures, since a conflation method may produce very different results depending on the measures used (Xavier et al., 2016). Numerous similarity measures have been developed in the literature based on the geometries, attributes, or spatial contexts of geographic features, with geometric and attribute-based similarity measures (Xavier et al., 2016) being the most commonly used methods.

A straightforward way of measuring the geometric similarity of two features is to compute a certain geometric property for each feature and compare the difference in that geometric property. For example, Zhang et al. (2012) used geometric properties including size, shape, and orientation to match polygons of building footprints. Each

geometric property is described by a number (or index). This includes a shape index based on well-known compactness measures (MacEachren, 1985; Wentz, 1997), an orientation index based on a statistical weighting method (Duchêne et al., 2003; Zhang et al., 2012), and the size of the polygon itself. For each index, the dissimilarity between two building polygons is then computed as either the difference or the ratio between their values of the index. Similarly, Yang, Zhang, and Luan (2013) have used the length and orientation of linear features to match roadways. Other scholars have used geometric properties including size (Tang, Hao, Zhao, & Li, 2008), bounding box (Tong, Shi, & Deng, 2009), and inertial axis to compare geographic features.

A potential limitation of geometric property-based methods is that the geometry of each feature is reduced to a single number on an individual basis, and specific information about the geometry may be lost before the comparison. For example, Wentz (1997) found that geometries with very different shapes can have near-identical compactness measures. Researchers (Tang et al., 2008; Yang et al., 2013; Zhang et al., 2012) often find it necessary to use other similarity metrics such as positional similarity to support the matching process (Xavier et al., 2016).

Similarity can also be computed directly from the coordinates of two geometries. For example, the similarity between two areal features can be computed based on their percentage overlap (Ruiz-Lendínez, Ariza-López, & Ureña-Cámara, 2013). This polygon overlap method may also be used to measure the similarity of two linear features if their buffer polygons are generated first and then compared. This leads to the simple buffer method (SBM) of Goodchild and Hunter (1997). In a similar vein, one can employ various distance metrics between features to measure similarity. The simplest example is the Euclidean distance, which has been widely used to match point features (Beeri, Kanza, Safra, & Sagiv, 2004). More complex distance metrics are required to match linear and areal features. Considering the relevance to the proposed method, in the sequel, we introduce three types of distance measures: point-offset distance, shape-based, and orientation-based distances.

2.1.1 | Point-offset distance measures

A type of generic geometric similarity for points, lines, and polygons is pointwise offset-based distances such as the widely used Hausdorff distance (Chehreghan & Abbaspour, 2017b; Li & Goodchild, 2010, 2011; Tong et al., 2014; Xavier et al., 2016). The *directed* Hausdorff distance $H_d(A,B)$ from a feature A to a second feature B is:

$$H_d(A,B) = \max_{p \in A} d(p,B)$$

where $d(p,B) = \min_{q \in B} d(p,q)$ is the straight-line distance from any given point p of A to feature B. $H_d(A,B)$ is the maximum of all such distances from feature A. The *directed* Hausdorff distance reflects the maximum amount of pointwise offset from one feature to another.

Pointwise distance measures such as the Hausdorff or Frechet distance may be incomplete. If a linear feature is very short, it may be matched to many neighboring features with very different shapes and orientations by the Hausdorff distance. Figure 1 presents an example of matching roads in two street networks near Alameda Avenue, Santa Barbara, CA. The two street networks are from OSM (in green) and the U.S. Census TIGER/Line (red), respectively. Figure 1 demonstrates the matches between roads using the offset-based directed Hausdorff distance and a conflation model to be discussed in Section 3. Each arrow represents a partial match—that is, the source feature corresponds to a part of the target feature. We can observe many partial matches near Alameda Avenue here because it was divided into many small segments in the OSM data (green), but it was represented as one polyline in the TIGER/Line data (red).

Moreover, we can observe that a fire lane in the OSM network is matched to Alameda Avenue in the TIGER/Line network. Clearly, this match is incorrect (rendered as a yellow arrow in Figure 1) as the fire lane has very different shape (and orientation) to Alameda Avenue. Yet an offset-based similarity metric such as the Hausdorff

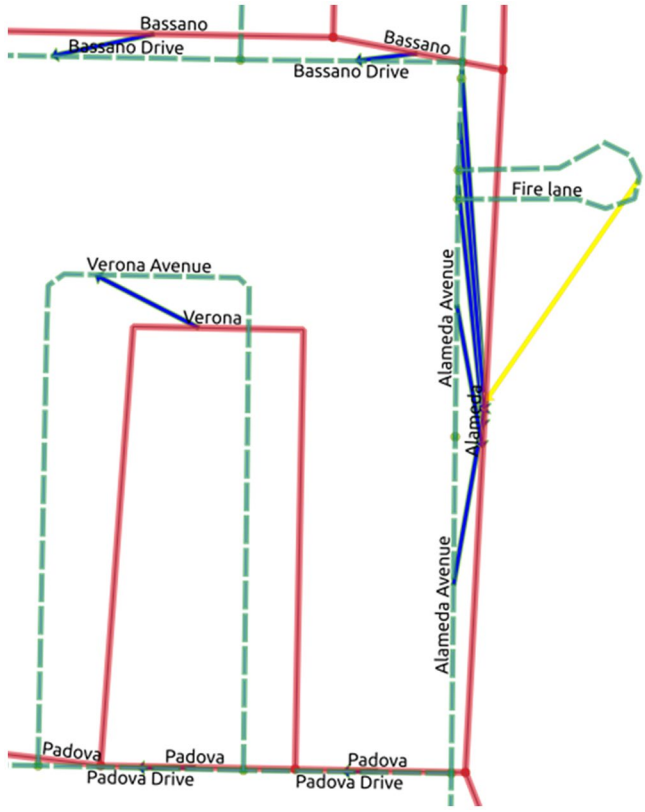

FIGURE 1 An incorrect match (yellow arrow) of roads caused by insensitivity to shape and direction using Hausdorff distances. Blue arrows represent correct matches

distance would not be able to tell, because the fire lane is a small feature with a total length of about 150 m (or effectively 70 m considering the fact that it is almost a double line). The directed Hausdorff distance from the fire lane to Alameda Avenue will be 70 m at most, which is comparable to the normal positional displacement between real corresponding features (such as the polylines representing Verona Avenue nearby).

Figure 2a presents the specific computation of the directed Hausdorff distance from the geometry of the fire lane to that of Alameda Avenue in Figure 1. The directed Hausdorff distance (53.66 m) is the distance from the dead-end loop of the fire lane to its nearest point on Alameda Avenue. Figure 2b presents the directed Hausdorff distance in the opposite direction, from Alameda Avenue to the fire lane. We can observe that the directed Hausdorff distance in this direction is much larger (216.97 m). The large directed-distance value in Figure 2b indicates correctly that the geometry of Alameda Avenue is unlikely to correspond to a part of the fire lane. The relatively small distance value in Figure 2a, however, is misleading. One may be led to believe that the fire lane corresponds to a part of Alameda Avenue, even though they look different. Clearly, the difference in shape and orientation here should be accounted for during the matching process.

In addition to the Hausdorff distance, the Frechet distance (Alt & Godau, 1995; Eiter & Mannila, 1994) is also based on pointwise offset. Also known as the dog-leashing distance, it is defined to be the minimum length of a leash that allows a dog and its owner to walk along their respective paths without backtracking (Chambers et al., 2010). The Frechet distance is better at differentiating certain circuitous curves but is also more expensive computationally. It has been used in GIS for matching coast lines (Mascret, Devogele, Le Berre, & Hénaff, 2006) and other linear features (Devogele, 2002).

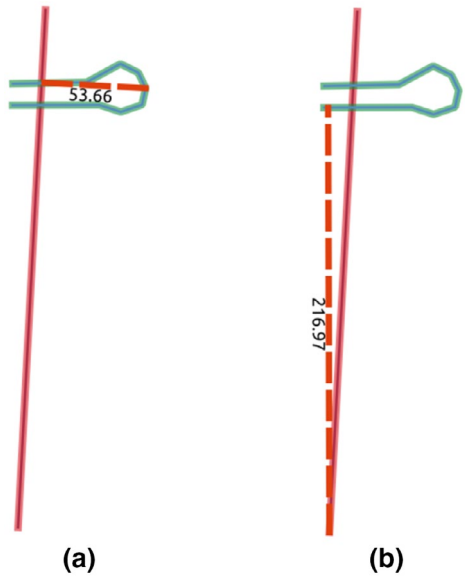

FIGURE 2 Directed Hausdorff distances between the fire lane (OSM) and Alameda Avenue (TIGER) in Figure 1: (a) from OSM to TIGER; and (b) from TIGER to OSM

In solving the gateway shortest-path problem, researchers have studied the similarity of alternative paths (Lombard & Church, 1993; Matisziw & Demir, 2016) between a fixed origin–destination pair in one network dataset. The spatial difference between two paths is measured either as the total area separating the two paths or as two paths' average area of deviation between their unshared portions. Two paths may share certain network nodes, and the areal deviation in these shared portions is counted as zero.

These methods address a different problem from conflation. While conflation aims to match *two* heterogeneous datasets, the above methods are aimed at finding the deviation of paths in *one* dataset. In conflation, counterpart polylines from two different data sources rarely have any shared portion. However, the areal deviation metrics are similar to the Hausdorff and Frechet distances in that they all measure deviation. Their difference is that the areal deviation metrics measure the accumulated or average deviation while the Hausdorff distance or Frechet distance measures the worst-case deviation. Neither the Hausdorff distance nor the areal deviation metrics consider angular differences as we do in this article. Therefore, they may be insensitive to shape differences in Figure 2a. In addition, the areal deviation metrics are defined between a fixed pair of origin and destination and imply a one-to-one comparison between two lines. Partial matching is therefore not accounted for.

2.1.2 | Shape measures

Shape is another criterion that can be used to tell the difference between linear features. The standard TF distance (Arkin et al., 1991) was designed to compare shapes of two polygons that may be in different orientations and scales. Frank and Ester (2006) used the TF distance to evaluate the quality of generalized maps. Li, Li, and Xie (2017) used the TF distance to evaluate the difference between different generalizations of a building's footprint that they produced using a "morphing" algorithm. The possibility of using the TF distance to measure polyline similarity was also discussed by Zhang (2009) and Chehreghan and Abbaspour (2017a), but no experimental results are reported.

The turning function is a function for describing the shapes of objects using cumulative turning angles. In the function, the angles between the segments of a polygon's boundary and the horizontal axis are accumulated over the length of the boundary lines. The lengths of the two boundary lines are typically scaled to 1 (unit length) and

the area between the cumulative functions of two lines is defined to be their shape dissimilarity. The area is computed as the accumulated L_p distance of the TF over the unit length, with L_2 distance being used originally (Arkin et al., 1991). Zhang (2009) and Chehreghan and Abbaspour (2017a) described the formula of the shape dissimilarity measure between polylines A and B (using L_1 distance) as shown in Equation (1):

$$TF(A,B) = \int_0^1 f\left(\theta_A(t), \theta_B(t)\right) dt$$
$$f\left(\theta_A, \theta_B\right) = \begin{cases} |\theta_A - \theta_B|, \text{if } |\theta_A - \theta_B| \le \pi \\ 2\pi - |\theta_A - \theta_B|, \text{otherwise} \end{cases} \quad (1)$$

In the above, $\theta_A(s)$ is the turning function of polyline A; t is the position in percentage length for a point on A measured from the starting point of A. Of note is that this definition assumes that the two polylines to be compared are approximately of the same length, and are scaled to unit length. The comparison of turning angles is therefore performed on the unit interval [0, 1]. This assumption may not hold if one polyline is a (proper) part of the other (or in a partial match situation in Figure 1). As will be discussed in Section 3, we allow the features with different lengths to be compared in the proposed MTFD distance, and overcome the scale-insensitivity by not rescaling the line features at all.

2.1.3 | Orientation measures

Orientation is another useful geometric property for measuring the similarity between polylines (Zhang, 2009) and polygons (Duchêne et al., 2003; Zhang et al., 2012). According to Zhang (2009), the orientation difference between two polylines A and B can be calculated by

$$d_{\text{Orient}}(A,B) = |\alpha_A - \alpha_B| = \arccos\left(\frac{v_A \cdot v_B}{|v_A| \cdot |v_B|}\right) \quad (2)$$

where α_A, α_B are orientations of A and B; v_A is the vector from the starting point of polyline A to its end point; and v_B is a similar vector for polyline B. This definition of orientation difference is effective when two polylines are relatively straight. However, when they are curved, this definition will be problematic as it does not reflect variation of directions within each polyline. Moreover, this definition may not be suitable for the partial matching of a shorter polyline to a longer one to which it "belongs" (Figure 1). As we will demonstrate in Section 3, a more subtle definition of orientation difference (or rotation) is needed for the partial matching of polylines. We will show that one can measure the shape and orientation difference in one algorithm.

In addition to geometric similarity metrics, attribute metrics compute the difference of two features based on their non-spatial properties, such as street names or place names (McKenzie, Janowicz, & Adams, 2014). For example, the Levenshtein distance measures the difference between two strings as the number of operations (insertion, deletion, or substitution) required to change one string into the other. Other string distances, such as the Hamming distance, have been used to match streets (Li & Goodchild, 2011). While the focus of this article is on enhancing geometric similarity measures, it should be noted that attribute-based similarity measures can be incorporated, when available, to enhance a geometric similarity measure.

2.2 | Conflation methods

Given one of the similarity measures mentioned above, a method is needed to decide which pairs of features should be selected as matched features. One of the earliest match-selection methods is the so-called one-sided nearest-neighbor join (Beeri et al., 2004). In its simplest form, one can match a feature in one dataset to its closest feature in the other dataset with respect to a distance or similarity metric. This kind of operation can readily be performed using many existing GIS packages or spatial databases.

However, the one-sided nearest-neighbor join is flawed in that the closeness relation can be inconsistent. As pointed out by Beeri et al. (2004) (illustrated in Figures 3a,b), if feature *A* in dataset 1 has the smallest distance to feature *B* in dataset 2, feature *B* can have the smallest distance to an entirely different feature *C* in dataset 1. Figure 3a presents an example of comparing TeleAtlas road data (green) and TIGER data (red) in Santa Barbara, CA. In the figure, Calle Grananda street in TIGER is the closest to the same road in the TeleAtlas data. However, Calle Grananda street in TeleAtlas is the closest to Colorado street in TIGER. Clearly, if one performs two nearest-neighbor joins starting from dataset 1 and dataset 2, respectively, one may draw different conclusions about which feature one should match *B* in Figure 3b (or Calle Granada in TeleAtlas in Figure 3a) to.

One possible method to avoid the aforementioned inconsistency is to use a "greedy" strategy called the *k*-closest pair query (KCPQ) [e.g., Ahmadi & Nascimento (2016) and Equation (1)]. It is widely used in the database literature and in pattern recognition due to its simplicity. The KCPQ method is iterative. In each iteration, it selects the closest pair of features from the two datasets as a pair of matched features, and then removes them from both datasets. The selection process stops after *k* pairs of features are selected or when no more features can be matched. Since a matched pair is excluded from further consideration, there will be no conflict in the selected matches. By design, KCPQ can only capture one-to-one matches. That is, a feature in a dataset can be matched to at most one feature in the other dataset. Due to its widespread use and simplicity, KCPQ will be used as one of the match-selection methods in this article to test the proposed similarity measures.

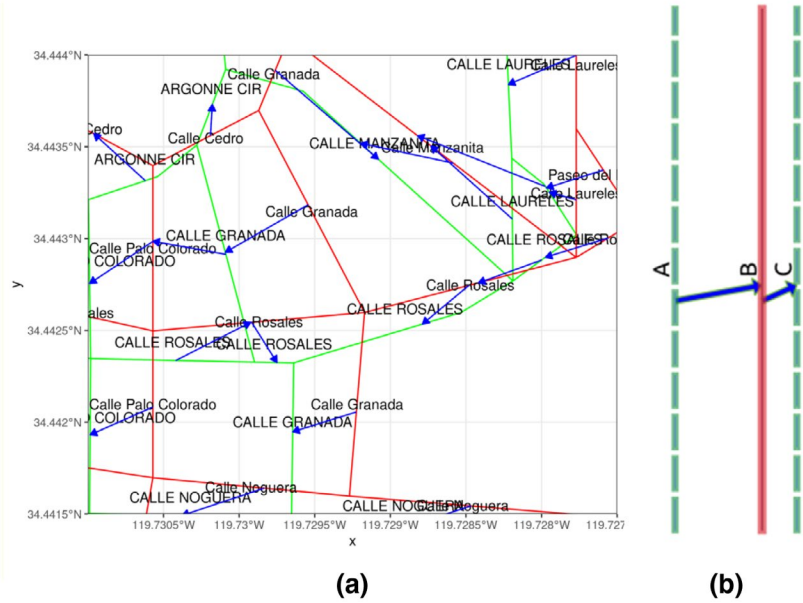

FIGURE 3 Inconsistent assignments in two directions of matching between datasets 1 (green) and 2 (red): (a) an example near Colorado St. in Santa Barbara, CA for TeleAtlas (green) versus TIGER (red) road networks; and (b) a simplified example

Beeri et al. (2004) attempted to solve the inconsistency between two opposite nearest-neighbor joins by computing a more general confidence value for matching features in what they called a probabilistic method. The confidence value for a pair of features is proportional to their similarity measure. A pair of features is matched if their confidence value is greater than a threshold value. With a similarity measure akin to the directed Hausdorff distance, Tong et al. (2009) extended the work of Beeri et al. (2004) so as to allow partial matching and handle one-to-many matches.

Another group of matching methods is the optimization-based methods, in which the feature-matching problem is formulated as an optimization problem of minimizing the total dissimilarity between matched features. Li and Goodchild (2010) formulated the feature-matching problem as an assignment problem, a classic model in operations research for crew assignment. It aims to assign a set of n workers to the same number of jobs, assuming that each worker i has a different cost c_{ij} for completing job j. The assignment problem minimizes the total assignment cost under the constraint that each worker is assigned to exactly one job. The assignment problem can be used for conflation by treating the features of one dataset as the workers and those of the other dataset as the jobs. The dissimilarity between a pair of features is then used as the assignment cost. Similar to the KCPQ method, the assignment problem is suitable for one-to-one matches only.

Multiple attempts have been made to extend optimization methods to handle (two-sided) one-to-many matches. Li and Goodchild (2011) developed a variant of the assignment problem that utilizes the directed Hausdorff distance to capture part–whole relationships. They employed two independent assignment-like models, one for each direction of matching, to find one-to-many matches in both directions. They then resolved the conflicts described by Beeri et al. (2004) (see Figure 3) after the optimization by deleting inconsistent matches. Tong et al. (2014) developed a hybrid model based on applying the assignment problem and then a heuristic method (logistic regression) in tandem. More recently, Lei, Church & Lei (In review) proposed a unified conflation model (in a companion article), which solves the consistency issue of opposite assignments (Beeri et al., 2004) (Figure 3) during optimization. Rather than removing inconsistent matches afterwards, they developed structural constraints to forbid conflicting assignments in advance and unify matches in opposite directions in one model.

Beyond one-to-many conflation methods, some methods can handle the more complex "many-to-many" matches. Walter and Fritsch (1999) proposed a "buffer-growing process," which connects adjacent polylines to form paths and then matches the paths. Masuyama (2006) proposed a similar method for matching polygons (census boundaries of 1990 and 2000, respectively), in which the author manually merged smaller boundary polygons into larger polygons until one-to-one matches can be established.

We have reviewed closely related methods in the literature both for measuring feature similarity and for selecting matches. This review is not comprehensive by any means. The reader may refer to Xavier, Ariza-López, and Ureña-Cámara (2017) for more comprehensive reviews of similarity measures, as well as Ruiz et al. (2011) for reviews of various processes of conflation, including useful pre- and post-processing procedures.

3 | METHOD

This section presents the definition of the proposed MTFD, which measures the difference of linear features both in terms of shape and orientation. We then demonstrate how to use the new metric in combination with traditional point-offset-based metrics to match linear features.

3.1 | Limitations of the standard TF distance

The TF distance (Arkin et al., 1991) was designed originally in the image processing and pattern recognition literature for matching the shapes of the same object appearing in different image scenes. It describes polygon

boundaries using the mathematical construct of TFs and measures the difference between the shapes of two polygon boundaries as the difference of their TFs. The TF distance is rotation- and scale-invariant. This is desirable because the rotation and scale difference between objects in different image scenes is often caused by a difference in perspectives and does not matter. Due to its generality, the TF distance has found a wide range of applications in matching objects in computer vision and many other fields, including GIS.

However, the standard TF distance has a number of restrictions that limit its usage in conflating geographic features in maps and GIS. First, rotation is important in map data. Large discrepancies of orientation between corresponding features (i.e., rotation) rarely happen in proper maps. Rotation should not be ignored in the conflation of maps. For example, two straight, perpendicular road segments have zero shape distance, but they are unlikely to represent the same road in two maps of the same coordinate system.

Second, the standard TF distance cannot handle partial matches. Instead, it can only match an entire polygon to another one. For example, a polygon A will match perfectly with itself. But if one splits A into two equal halves, A_1 and A_2, neither part will match A due to the deviation of turning angles at the common boundary of A_1 and A_2. This means that the standard TF distance can only handle one-to-one matches; it cannot handle the more complex many-to-one (or many-to-many) matches because a part of the shape can have a very large distance to the shape itself. This is where the TF distance falls short while quasi-distance measures such as the directed Hausdorff distance (or the proposed partial TF distance) work well.

Third, and related to the one-to-one matching assumption, the standard TF distance depends on rescaling two shapes so that their boundaries have the same length (or unit length). While scale-invariance is required in object recognition, it is undesirable to match two features of different sizes in GIS, since two shapes representing the same object should have approximately the same size. Therefore, introducing scale-invariance can lead to false matches.

Fourth, the standard TF definition is mathematically dependent on the fact that the boundaries of the two polygons form closed loops, and the extension of turning functions cyclically to an infinite domain. Such an assumption does not apply to non-closed polylines. Nonetheless, the TF is a powerful tool in matching shapes in that it provides a direct measurement of the shape deformation between two counterpart features. This is often lacking in other similarity measures such as the Hausdorff distance or the Frechet distance. Next, we present the orientation-sensitive MTFD, based on extending the standard TF distance (Arkin et al., 1991).

3.2 | MTFD

To define the MTFD, the following notation is needed. Given a polyline A (or a curve in general), suppose a person walks along it from the beginning to the end at unit speed (e.g., 1 m/s). Then each point on A is uniquely identified by the time t at which it is visited. Therefore, we can write each point of A as $A(t)$. Since we assume a unit speed of travel, t also represents the *length* of curve on A between the starting point and $A(t)$. Essentially, t defines an *intrinsic coordinate system* for points on the curve, which is commonly known as a *linear referencing* system in GIS. The use of the intrinsic coordinate t to uniquely identify any point $A(t)$ of A is also called a *parameterization* of A, where the time t is also referred to as a parameter of the curve A.

Mathematically, the turning function $\Theta_A(t)$ of a curve A is a function of the tangent direction at any point $A(t)$ on A versus intrinsic coordinate t (Arkin et al., 1991). The tangent direction at $A(t)$ is also called the *turning angle*, and is defined in radians and relative to the x-axis. Figure 4 depicts the turning angles and turning function of the fire lane in Figure 1, respectively. As shown in Figure 4a, the turning angle for a polyline is constant within each of its straight-line segments. Therefore, its turning function (Figure 4b) is a step function, in which changes of direction angle only happen at each time point t associated with a vertex of the polyline. If a polyline is a straight line (similar to Alameda Avenue in Figure 1), its turning function will be a constant function. Clearly, the difference between the turning functions of the fire lane and Alameda Avenue (Figure 1) will be large. The turning function

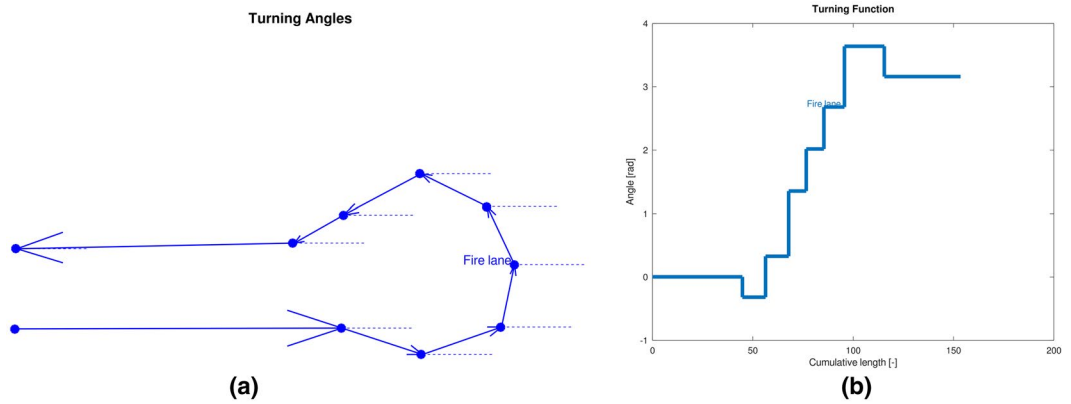

FIGURE 4 Definition of the TF of a polyline: (a) polyline of the fire lane in Figure 1; and (b) turning function of the fire lane

of a polyline A essentially carries the same information as the curvature at point $A(t)$, as the curvature is merely the rate of change of the turning function at t. Roughly speaking, the proposed MTFD is used to gauge the difference between the two turning functions.

Given the above definitions, the basic assumptions of the MTFD are as follows. First, it is intended for matching a linear feature A to a larger linear feature B, and A possibly corresponds to a part of feature B in reality. We call A the "from-feature" and B the "to-feature." Suppose the lengths of A and B are l and L, respectively, with $l \le L$. Then in case of a perfect match, all points of A (i.e., $A(t), t \in [0, l]$) should coincide with a subset of B starting at a certain point $B(s)$ (i.e., $A(t) = B(t+s), t \in [0, l]$). In this case, the directions (and locations) of A and B must match everywhere in this interval of B, and their MTFD (as well as Hausdorff distance) should be zero. The MTFD distance is non-zero otherwise. In the remainder of the article, we call the time s associated with the starting point $B(s)$ the *starting position* of comparison. Each starting position s on B may give rise to a different value of dissimilarity. The MTFD is defined to be the smallest dissimilarity of all starting positions s.

Intuitively speaking, we restrict our attention to the part–whole or many-to-one matches. This is more general than the one-to-one matches considered in the standard TF distance. In addition, we do not normalize the two polylines to the same length as in the traditional TF distance, and instead measure dissimilarity without any rescaling. This makes our similarity measure sensitive to differences in scale, a property that is essential in comparing GIS features. We do not consider the symmetric case of matching a larger feature B to a smaller feature A. This is because without rescaling, a portion of B will have no counterpart in A, and it is not obvious how a distance or similarity metric for this part of B can be defined.

Second, if rotation is required to "align" feature A to feature B (or in other words, a systematic difference of orientation exists between the two features), the rotation is considered a type of dissimilarity between the two features. Just like shape deformation, rotation is an angular measurement. Among all possible starting positions of comparison, the position that minimizes the *sum* of shape deformation and rotation is used to characterize the difference between the features.

Third, we rely on well-established metrics (such as the Hausdorff distance) to determine the amount of pointwise offset between two features. We then combine a point-offset metric with our angular metric (deformation and rotation) to form a composite similarity metric.

Figure 5 presents an example of the turning functions of two polylines A (blue) and B (red). Under the aforementioned assumptions, we can define the MTF distance between two polylines as follows. Similar to Arkin et al. (1991), we measure the angular difference of polyline A with respect to polyline B using the L_p distance

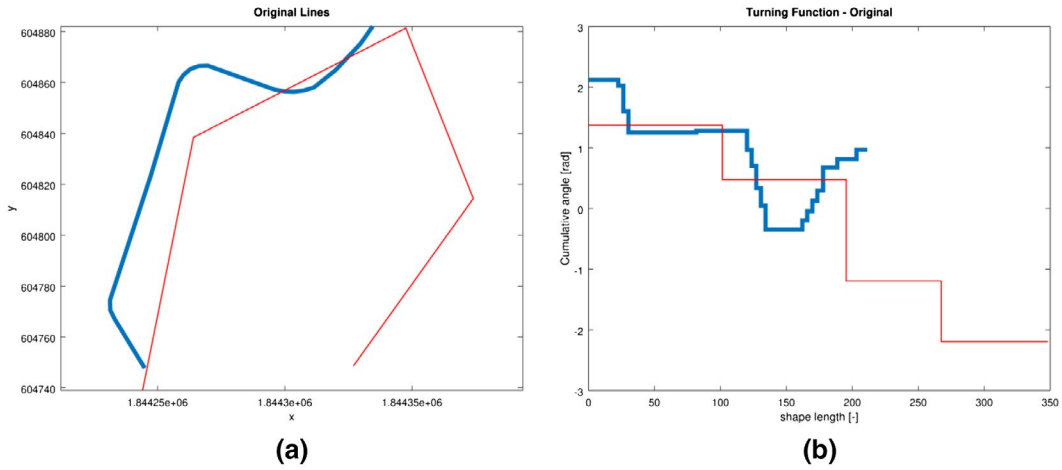

FIGURE 5 Example directed polyline-based turning functions: (a) two polylines; and (b) their turning functions

(with $p = 2$) between their turning functions in the space of the intrinsic coordinate (or time) t (Figure 5b). For a given starting position s on B, we measure the angular difference between $A(t)$ and $B(t + s)$ for each value of time $t \in [0, l]$, which is $\Theta_A(t) - \Theta_B(t + s)$. Conceptually, the accumulated angular difference between $A(t)$ and $B(t + s)$, $t \in [0, l]$ contains two types of difference—a global rotation between A and B, and the shape deformation (i.e., the remaining difference after the global rotation is removed). To distinguish these two types of difference, we create an auxiliary variable θ for the global rotation, and the correct value of θ is yet to be found. Then, $\Theta_A(t) - \Theta_B(t + s)$ decomposed into two parts: the overall rotation θ and the residue $\Theta_A(t) - \Theta_B(t + s) - \theta$, which represents the *shape deformation* (called *deformation* hereafter). The deformation between A and B for a given starting position s and rotation θ is:

$$m_p(A, B, s, \theta) = \frac{\left(\int_0^l \left|\Theta_B(t + s) - \Theta_A(t) + \theta\right|^p dt\right)^{1/p}}{\sqrt{l}} \tag{3}$$

where l and L are the lengths of A and B, respectively, with $l \leq L$. Since shape deformation should not include any angular difference that can be accounted for by the systematic rotation, we need to find the minimum $m_p(A, B, s, \theta)$ over θ. That is we seek, for any given s, the following:

$$\beta(A, B, s) = \min_{\theta} m_p(A, B, s, \theta) \tag{4}$$

as well as the associated optimal θ value in Equation (4):

$$\alpha(s) = \theta_s^* \tag{5}$$

The optimal value $\alpha(s)$ represents the estimated rotation, and $\beta(A, B, s)$ is the deformation. Then, at a second (higher) level of optimization, we seek to find the starting position s that minimizes the weighted sum of deformation and rotation as follows:

$$m(A, B) = \min_{s} \left(\beta(A, B, s)^2 + w_r \cdot \alpha(s)^2\right)^{1/2} \tag{6}$$

In the above, we use a weight value w_r to represent the relative importance of seeking a position that minimizes rotation versus one that minimizes shape deformation. $m(A, B)$ is the MTFD. If the weight value w_r is zero and $l = L$, then the MTFD reduces to the regular TF distance. The solution of the lower-level minimization problem (4) can be found by setting $\frac{\partial m_p(A, B, s, \theta)}{\partial \theta} = 0$, which renders:

$$a(s) = \theta_s^* = \frac{1}{l}\int_0^l \Theta_A(t) - \Theta_B(t+s)\,dt \tag{7}$$

Insert Equation (7) into Equation (4) and, after simplification, we obtain the estimated deformation:

$$\begin{aligned}\beta(A,B,s) &= \frac{1}{\sqrt{l}}\left(\int_0^l \left(\Theta_B(t+s) - \Theta_A(t) + \alpha(s)\right)^2 dt\right)^{1/2} \\ &= \frac{1}{\sqrt{l}}\min_s\left(\int_0^l \left(\Theta_B(t+s) - \Theta_A(t)\right)^2 dt + l\cdot\alpha^2(s) - 2\alpha(s)\cdot\int_0^l \Theta_A(t) - \Theta_B(t+s)\,dt\right)^{1/2} \\ &= \frac{1}{\sqrt{l}}\min_s\left(\int_0^l \left(\Theta_B(t+s) - \Theta_A(t)\right)^2 dt - l\cdot\alpha^2(s)\right)^{1/2}\end{aligned} \tag{8}$$

The overall MTFD (6) can be obtained by substituting Equations (7) and (8) into Equation (6):

$$\begin{aligned}m(A,B) &= \frac{1}{\sqrt{l}}\min_s\left(\int_0^l \left(\Theta_B(t+s) - \Theta_A(t) + \alpha(s)\right)^2 dt + w_r\cdot l\cdot(a(s))^2\right)^{\frac{1}{2}} \\ &= \min_s\left(\frac{1}{l}\cdot\int_0^l \left(\Theta_B(t+s) - \Theta_A(t)\right)^2 dt + \left(w_r - 1\right)\cdot\alpha^2(s)\right)^{1/2}\end{aligned} \tag{9}$$

In addition, we define $m(B, A) = \infty$.

3.3 | Choosing minimum shape difference versus minimum rotation

The above computation implicitly looks for the portion of the to-feature B from a certain point $B(s)$ that best matches the from-feature A. We seek to find the starting position s of B with the minimum overall deformation and rotation. Our *criterion* for aligning A with B (in the space of time t) is more general than that of the standard TF distance. The new criterion is needed because we may get wrong answers about the relationship between two features if we use only deformation.

Figure 6a presents an example of a potential partial match of a straight line A to a polyline B with three segments. If they are in a map, one would expect that A should be matched to the two horizontal segments of B in Figure 6a, since it will incur a reasonably small shape deformation, but nearly no rotation. However, if deformation is the only match criterion as with the standard TF metric, A will be matched to the longer segment of B that is almost perpendicular to A, because this match will incur zero shape deformation (as shown in Figure 6b). Meanwhile, the TF distance will report that A is perpendicular to B. This is clearly not the case in GIS. To avoid such issues, we set the weight value of the rotation objective in Equation (6) to a positive value. In this article, we assume that $w_r = 1$.

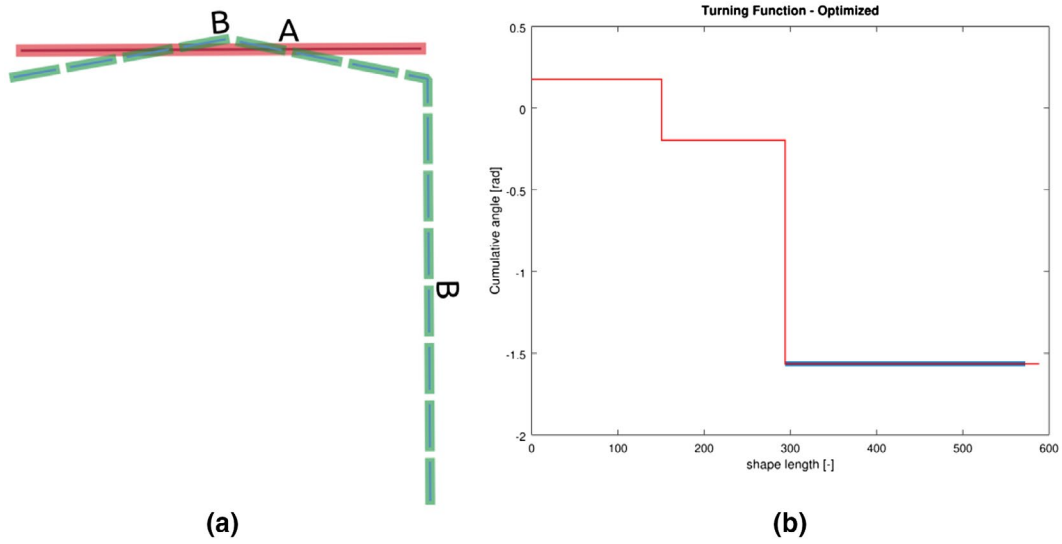

FIGURE 6 Minimizing shape deformation is insufficient: (a) potential partial match; and (b) optimal match by shape alone

3.4 | Finite-optimality set for locating partial matches

The work of Arkin et al. (1991) is important not only because they defined the TF distance mathematically, but also because they identified a *finite* set of starting positions on B called the critical events. They proved, in a series of lemmas and theorems, that it is only necessary to evaluate shape deformation at this finite set to find the optimal match position s. In this subsection, we provide a proof of the finite optimality of the generalized MTFD distance (6) and show that a similar finite-optimality set is sufficient for the new distance.

Definition 1 Given a polyline A, a time point t_0 is called a vertex position of A if $A(t_0)$ is a vertex of A.

Definition 2 Given a polyline B, we call the polyline $B_s(t) := B(t+s)$ the shifted version of B.

Remark 1 By Definition 1 and Definition 2, if t_0 is a vertex position of $B_{s'}(t)$, then $t_0 - s + s'$ is a vertex position of $B_s(t)$.

Definition 3 The starting position s is called a crossing point for polylines A and B if there exists a time $t \in [0, l]$ such that both $A(t)$ and $B(t+s)$ are vertices.

Intuitively, a starting position s is a crossing point if two persons walk on A and B at the same speed (unit speed), starting from $A(0)$ and $B(s)$, respectively, and encounter a pair of vertices of A and B at the same time. For example, the first crossing point is $e_1 = 0$ (because with $s = 0$, the two persons arrive at the starting vertices of A and B at the same time, $t = 0$), and the last crossing point is $e_K = L - l$ for some integer K (at which the agents arrive at the end vertices of A and B at the same time, $t = l$). Similarly, all s for which $B(s)$ are vertices are crossing points because vertices $A(0)$ and $B(s)$ are encountered at the same time, $t = 0$. In general, if A and B have m and n vertices, respectively, there are at most $O(m \cdot n)$ crossing points because each vertex of B can coincide with at most m vertices of A.

Lemma 1 *Let $e_1, \ldots, e_K$ be the set of crossing points. For any interval (e_i, e_{i+1}), $\alpha(s)$ is a linear function of the form:*

$$\alpha(s) = g_i \cdot s + h_i$$

for some constants g_i and h_i, and the objective function $\beta^2(s)$ defined in Equation (8) is a quadratic function in (e_i, e_{i+1}).

Proof We will first show that for a given starting position s on polyline B, the difference in the turning angle $\Theta_A(t) - \Theta_B(t+s)$ is a piecewise constant function in t. Consider the following intervals $I_i = [t_{i-1}, t_i), i = 1, \ldots, q$, where the set of $\{t_i, i = 1, \ldots, q\}$ with $t_i < t_{i+1}$ includes all the vertex positions of A (Definition 1) and B_s (Definition 2). In particular, we have $t_0 = 0, t_q = \min(l, L - s)$. By Equation (7), the estimated rotation $\alpha(s)$ is:

$$\alpha(s) = \frac{1}{l}\sum_{i=1}^{q}\int_{t_{i-1}}^{t_i} \Theta_A(t) - \Theta_B(t+s)\,dt$$

For a fixed i, both $A(I_i)$ and $B(s + I_i)$ are straight-line segments since, by definition, the interior of I_i does not contain any vertex positions of A or B. Therefore, the difference in turning angles $\Theta_B(t+s) - \Theta_A(t), t \in I_i$ is a constant, which we denote as D_i. Then, the rotation $\alpha(s)$ in Equation (7) reduces to the following:

$$\alpha(s) = \frac{1}{l}\sum_{i=1}^{q} -D_i \cdot |I_i| \tag{10}$$

Next, we will show that $\alpha(s)$ changes linearly with s as long as s is not a crossing point. Suppose the starting position s is not a crossing point. When s is changed by an infinitesimal amount δs, let $\{\hat{t}_i, i = 1, \ldots, q\}$ with $\hat{t}_i < \hat{t}_{i+1}$ be the new set of vertex positions corresponding to A and $B_{s+\delta s}$, and $\hat{I}_i = [\hat{t}_{i-1}, \hat{t}_i)$. By a similar argument, we have:

$$\alpha(s + \delta s) = \frac{1}{l}\sum_{i=1}^{q} -D_i \cdot |\hat{I}_i| \tag{11}$$

We show next that the change in interval width $|\hat{I}_i| - |I_i|$ can have four possible cases, depending on what type the interval $|I_i|$ is:

1. If both t_{i-1} and t_i are vertex positions of B_s, by Remark 1, $t_{i-1} - \delta s$ and $t_i - \delta s$ are vertex positions of B_s. Therefore, by the definition of $\hat{t}_i$ and the fact that δs is infinitesimal, $\hat{t}_{i-1} = t_{i-1} - \delta s$ and $\hat{t}_i = t_i - \delta s$. Hence $|\hat{I}_i| - |I_i| = 0$.
2. If t_{i-1} is a vertex position of B_s, and t_i is a vertex position of A, then $\hat{t}_{i-1} = t_{i-1} - \delta s$ (due to the same reasoning as in 1) and $\hat{t}_i = t_i$ (due to the fact that vertex positions of polyline A do not change with s). Hence $|\hat{I}_i| - |I_i| = \delta s$.
3. If t_{i-1} is a vertex position of A and t_i is a vertex position of B_s, then $\hat{t}_{i-1} = t_{i-1}$ and $\hat{t}_i = t_i - \delta s$. Hence $|\hat{I}_i| - |I_i| = -\delta s$.
4. If both t_{i-1} and t_i are vertex positions of A, then $\hat{t}_{i-1} = t_{i-1}$ and $\hat{t}_i = t_i$. Hence $|\hat{I}_i| - |I_i| = 0$.

(Note that since s is not a crossing point, each t_i is either a vertex position of A or a vertex position of B_s, but not both. Hence the above cases cover all possible scenarios.)

Define an auxiliary vector $b = (b_1, \ldots, b_q)$, such that $b_i = 0$ if I_i is Type 1 or 4, $b_i = 1$ if I_i is Type 2, and $b_i = -1$ if I_i is Type 3. Then, $\alpha(s)$ changes by $\alpha(s + \delta s) - \alpha(s) = C \cdot \delta s$, where $C = \sum_{i=1}^{q} \frac{1}{l} b_i D_i$ is a constant. This implies that the derivative of $\alpha(s)$ is a constant, hence $\alpha(s)$ is linear. This linear relationship holds until s reaches a crossing point. This is because when s reaches a crossing point, some of the intervals I_i may shrink to zero width (i.e., disappear)

and new intervals with new angular difference between the two polylines may appear. On such occasions, new linear relationships between $\alpha(s)$ and s may take over. Therefore, $\alpha(s)$ is piecewise linear between each pair of consecutive crossing points. And we can write:

$$\alpha(s) = g_i \cdot s + h_i, s \in (e_i, e_{i+1}) \tag{12}$$

for some constants g_i and h_i. From the above computation, we have $g_i = \frac{1}{l}\sum_{j=1}^{q} b_j D_j$ and $h_i = \alpha(e_i) - g_i \cdot e_i = \frac{1}{l}\sum_{j=1}^{q}\int_{t_{j-1}}^{t_j} \Theta_A(t) - \Theta_B(t + e_i)\, dt - g_i \cdot e_i$.

By the same argument above, the first main term of $m(A, B)$ in Equation (9) (shown below) is a piecewise linear function of starting position s:

$$\frac{1}{l}\int_0^l \left(\Theta_B(t+s) - \Theta_A(t)\right)^2 dt = \frac{1}{l}\sum_{i=1}^{q} D_i^2 \cdot |I_i|$$

Since $\alpha(s)$ is piecewise linear, the second term $(w_r - 1) \cdot \alpha^2(s)$ is a piecewise quadratic in s and the boundaries of the pieces are in the set of crossing points. Therefore, $\beta^2(s)$ is a piecewise quadratic function of s in each interval (e_i, e_{i+1}).

Theorem 1 *If $w_r \le 1$, the minimizers of $m(A, B)$ in Equation (9) occur at the crossing points $E = \{e_i, i = 1, \ldots, K\}$. If $w_r > 1$, the minimizers occur either at the crossing points in E, or at the set of stationary points $F = \left\{f_i \,|\, f_i \in (e_i, e_{i+1}) \text{ and } f_i = \frac{\tilde{g}_i}{2(1-w_r)g_i^2} - \frac{h_i}{g_i}, i = 1, \ldots, K\right\}$, where $\tilde{g}_i = \frac{1}{l}\sum_{i=1}^{q} b_i D_i^2$.*

Proof $m(A, B)$ and $m^2(A, B)$ have the same minimizer, and $m^2(A, B)$ is piecewise quadratic in each (e_i, e_{i+1}) by Lemma 1.

If $w_r \le 1$, the fact that $w_r - 1 \le 0$ implies that the objective function in Equation (9) is concave (downwards). Hence, the minimizers only occur at $e_i, i = 1, \ldots, K$.

If $w_r > 1$, then $m(A, B)$ is convex (downwards) in (e_i, e_{i+1}), and the minimizer must be at either the boundary points e_i, or the stationary point determined by:

$$\frac{\partial\left(m^2(A, B)\right)}{\partial s} = \frac{1}{l}\sum_{i=1}^{q} b_i D_i^2 + 2w_r \alpha(s)\alpha'(s) = 0$$

Let $\tilde{g}_i = \frac{1}{l}\sum_{i=1}^{q} b_i D_i^2$, and inserting Equation (11) into the above, we have:

$$g_i \cdot s_i^* + h_i = \frac{\tilde{g}_i}{2(1 - w_r) g_i}$$

$$s_i^* = \frac{\tilde{g}_i}{2(1 - w_r) g_i^2} - \frac{h_i}{g_i}$$

Therefore, if $s_i^* \in (e_i, e_{i+1})$, *then* $f_i = s_i^*$ *is a possible minimizer of* $m(A, B)$.

The proof of the finite-optimality set above is similar, at least in spirit, to the proofs of the finite-optimality sets of the *p*-maxima problem (Church & Garfinkel, 1978) and the vector assignment ordered median problem (Lei & Church, 2015). In essence, we proved that improvements of the objective function (here, the sum of deformation and rotation) can be made in continuous regions of positions until certain critical event positions (*E* and *F* here) are encountered. Therefore, an optimal solution must exist among the break points. Overall, Theorem 1 means that to find the best position for matching *A* to *B*, we only need to *enumerate* the MTFD objective function at all critical events and then select the position with the smallest value of the following expression:

$$\frac{1}{l}\sum_i \left(D_i - \frac{\sum_j D_j |l_j|}{l}\right)^2 |l_i| + w_r \cdot \alpha(s)^2 = \frac{\sum_i D_i^2 |l_i|}{l} + (w_r - 1)\left(\frac{\sum_i D_i |l_i|}{l}\right)^2 \tag{13}$$

In particular, as we assume $w_r = 1$ in this article, we only need to find the smallest $\frac{\sum_i D_i^2 |l_i|}{l}$ for all the critical events along *B*.

3.5 | Combining angular difference and pointwise offset

To complement the angular discrepancy measure (shape deformation and rotation), we incorporate the directed Hausdorff distance and the MTFD to form a composite similarity measure that considers both pointwise offset and angular difference. The composite measure, called the directed turning function Hausdorff distance (DTFH distance), is defined as follows:

$$d_{DTFH}(A, B) = \begin{cases} \infty, \text{ if } \beta_{AB} > c_{shp}, \alpha_{AB} > c_{rot}, \text{ or } h_{AB} > c_{off} \\ m'(A, B) \cdot R \cdot w_a + h_{AB}, \text{ otherwise} \end{cases}$$

In the above, β_{AB} and α_{AB} are the shape deformation and rotation from *A* to *B*; h_{AB} is the directed Hausdorff distance (i.e., point "offset") from *A* to *B*. $m'(A, B)$ is a weak angular distance, defined as:

$$m'(A, B) = \min(m(A, B), m(B, A))$$

$m'(A, B)$ represents the minimum of the MTFDs (i.e., angular discrepancies) in the two opposite directions of matching. c_{shp}, c_{rot}, and c_{off} are cutoff values for shape difference, rotation, and offset, respectively. If any of them exceeds its associated cutoff value, the composite distance is defined to be infinity. Since the weak MTFD $m'(A, B)$ is in radians and the point-offset distance h_{AB} is in meters, we use a constant *R* to convert angular difference in radians to meters. In this article, we use $R = 100$ m/rad. This means that we equate 1 radian of angular difference (approximately 57°) to 100 m of point-offset distance. In addition, we use a relative weight value w_a on $\hat{m}(A, B)$ to represent the emphasis over angular difference versus positional difference.

We also define an overall cutoff value *c*. If $d_{DTFH} \geq c$, then we consider d_{DTFH} as infinity. The cutoff values should be chosen carefully to filter out only counterpart features that are unlikely to match. Of note is that we use the weak MTFD distance instead of the directed angular distance in the DTFH distance because the shape difference from the longer feature in *A*, *B* to the shorter one is not well-defined (and set to infinity).

It should also be noted that when the angular weight w_a is set to 0, the DTFH measure involves the Hausdorff distance measure with a preprocessing step using the multiple cutoff distances (e.g., rotation and shape deformation). So even if $w_a = 0$, the composite DTFH measure is not purely the Hausdorff distance measure.

3.6 | Choosing a match plan based on the combined metric and optimization

To evaluate the effectiveness of the DTFH distance, we use two simple selection methods mentioned in Section 2 in a set of experiments to match road features. The first is the KCPQ, widely used in the database literature. The KCPQ assumes a one-to-one correspondence between counterpart features, and therefore cannot handle many-to-one correspondence. To handle the many-to-one cases in part–whole relations, we use a unified conflation model developed in a companion article (Lei, Church & Lei, In review). It is an extension of the work of Li and Goodchild (2011), which uses two independent assignment problems (Li & Goodchild, 2010) to select part–whole matches in the two opposite directions of matching. By comparison, the unified conflation model harmonizes assignments in opposite directions in one model by utilizing new constraints to ensure the compatibility of opposite assignments. For the sake of completeness, we describe the unified conflation model below. The following notation is needed:

- i, I are the index and set of features in dataset 1
- j, J are the index and set of features in dataset 2
- $F = \{(i,j) \mid d_{TFH}(i,j) < c, i \in I, j \in J\}$ is the set of potential forward assignments for which assignment distances are less than the cutoff value
- $F_i = \{(i,j) \mid d_{TFH}(i,j) < c, j \in J\}$ is the subset of admissible forward assignments from feature $i \in I$, and $|F_i|$ is its size
- $B = \{(j,i) \mid d_{TFH}(j,i) < c, i \in I, j \in J\}$ is the set of backward assignments with below-cutoff distances
- $B_j = \{(j,i) \mid d_{TFH}(j,i) < c, i \in I\}$ is the subset of admissible backward assignments from feature $j \in J$, and $|B_j|$ is its size
- $c_{ij} = D + 1 - d_{TFH}(i,j)$ is a similarity measure between i and j, where D is the largest finite distance between features I and J. By definition, all c_{ij} and c_{ji} for F and B are positive.

The decision variables are:

- $u_{ij} = 1$ if feature $i \in I$ is matched to feature $j \in J$, or 0 otherwise. Semantically, $u_{ij} = 1$ means that feature i belongs to j, or i corresponds to a part of j.
- $v_{ji} = 1$ if feature $j \in J$ is matched to feature $i \in I$, or 0 otherwise. If both u_{ij} and v_{ji} are 1, i and j are considered to be the same feature.

Given this notation, the unified conflation model is:

$$\text{Maximize } Z = \sum_{(i,j) \in F} c_{ij} u_{ij} + \sum_{(j,i) \in B} c'_{ji} v_{ji} \tag{14}$$

Subject to:

$$\sum_{(i,j) \in F} u_{ij} \le 1 \quad \text{for each } i \tag{15}$$

$$\sum_{(j,i) \in B} v_{ji} \le 1 \quad \text{for each } j \tag{16}$$

$$\left(|B_j| - 1\right) u_{ij} + \sum_{k \in B_j, k \ne i} v_{jk} \le |B_j| - 1 \quad \text{for } (i,j) \in F \tag{17}$$

$$\left(|F_i| - 1\right) v_{ji} + \sum_{k \in F_i, k \ne j} u_{ik} \le |F_i| - 1 \quad \text{for } (i,j) \in B \tag{18}$$

$$u_{ij} \in \{0,1\} \quad \text{for each } (i,j) \in F \tag{19}$$

$$v_{ji} \in \{0,1\} \quad \text{for each } (j,i) \in B \tag{20}$$

The objective (14) of the unified conflation model is to maximize the total similarity between matched features. Constraint (15) maintains that a feature i in dataset I can belong to at most one feature in dataset J. Conversely, constraint (16) maintains the same condition as constraint (15) in the opposite direction. Constraint (17) ensures the compatibility of opposite assignments. In particular, it maintains that if i belongs to j (i.e., $u_{ij} = 1$), then j cannot belong to any feature k in I other than i itself (i.e., $\sum_{k \in B_j, k \neq i} v_{jk} = 0$). If i does not belong to j (i.e., $u_{ij} = 0$), then the constraint is not binding. Constraint (18) maintains compatibility of assignments in the opposite direction. Constraints (17) and (18) together prevent the inconsistency depicted in Figure 3 from happening. Constraints (19) and (20) define the assignment variables u_{ij}, v_{ji} as binary decision variables.

4 | EXPERIMENT

4.1 | Experiment settings

In this article, we have used road datasets (Figure 7) covering six test sites in Santa Barbara, CA as the test data. They have the same geographic extent as the datasets used by Li and Goodchild (2011), but come from different data sources. We use OSM and TIGER/Line datasets, both publicly available. The OSM dataset represents road networks of the study areas in January 2018. The TIGER/Line dataset is the same as the one used by Li and Goodchild (2011).

To evaluate the accuracy of the proposed method, we manually labeled matches for all test sites as ground-truth data. The accuracy is evaluated based on the widely used recall and precision rates, which compare the algorithm-predicted matches and the ground truth. Recall is defined as:

$$\text{Recall} = \text{TM}/\text{AM}$$

where TM is the number of true matches for which the algorithm and the ground truth agree, and AM is the number of all matches in the ground truth. Precision is defined as:

$$\text{Precision} = (\text{TM} + \text{TU}) / (\text{TM} + \text{TU} + \text{FM} + \text{FU})$$

where TU is the number of true unmatches (features that the algorithm correctly kept unmatched according to the ground truth), FM is the number of false matches (features that are falsely matched by the algorithm, but not matched in the ground truth), and FU is the number of false unmatches (features that are matched in the ground truth but falsely unmatched in the algorithmic result). Recall reflects the algorithm's capability in capturing true matches. Precision reflects additionally the algorithm's discriminative power to filter out false matches. To evaluate the average accuracy of the conflation methods, we also compute the F score from the recall and precision as follows:

$$F = 2 \cdot \frac{\text{Recall} \cdot \text{Precision}}{\text{Recall} + \text{Precision}}$$

We coded the TF-based distances in Matlab/Octave and computed the distance matrix for the metric between all pairs of features. As mentioned earlier, we use a default rotation weight (w_r) of 1.0. We then implemented the two conflation models. The first one is the KCPQ described earlier. The second is the unified conflation model

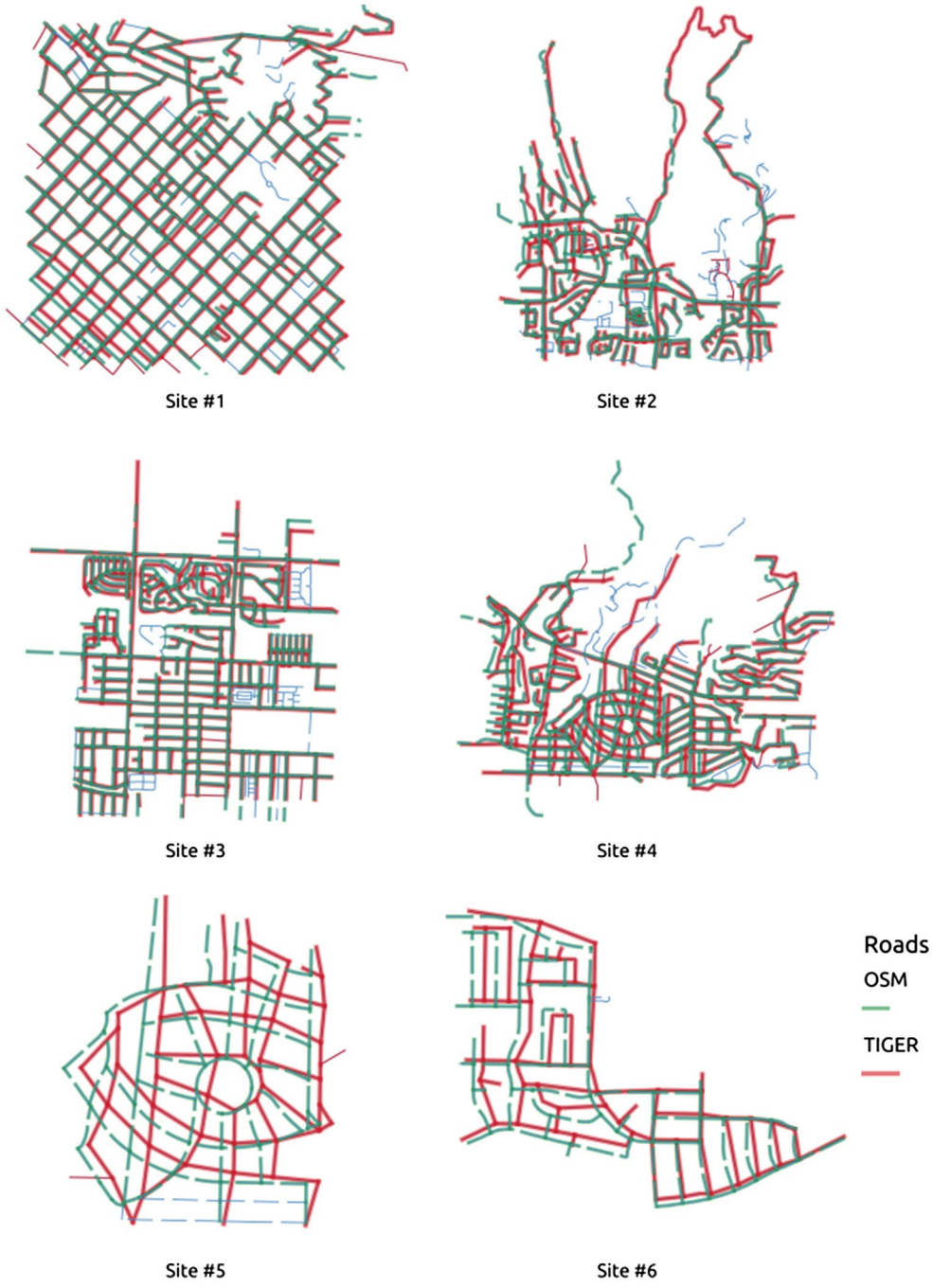

FIGURE 7 Road networks of six test sites in Santa Barbara, CA using OSM (green) and TIGER/Line (red). Matched features in the ground truth are depicted with thicker lines

(14)–(20), which is implemented as an integer linear program using IBM/ILOG CPLEX Studio 12.10. We implemented two versions of each model, with one version using only the directed Hausdorff distance and the other version using the hybrid DTFH distance. Both models are relatively straightforward to implement as they do not

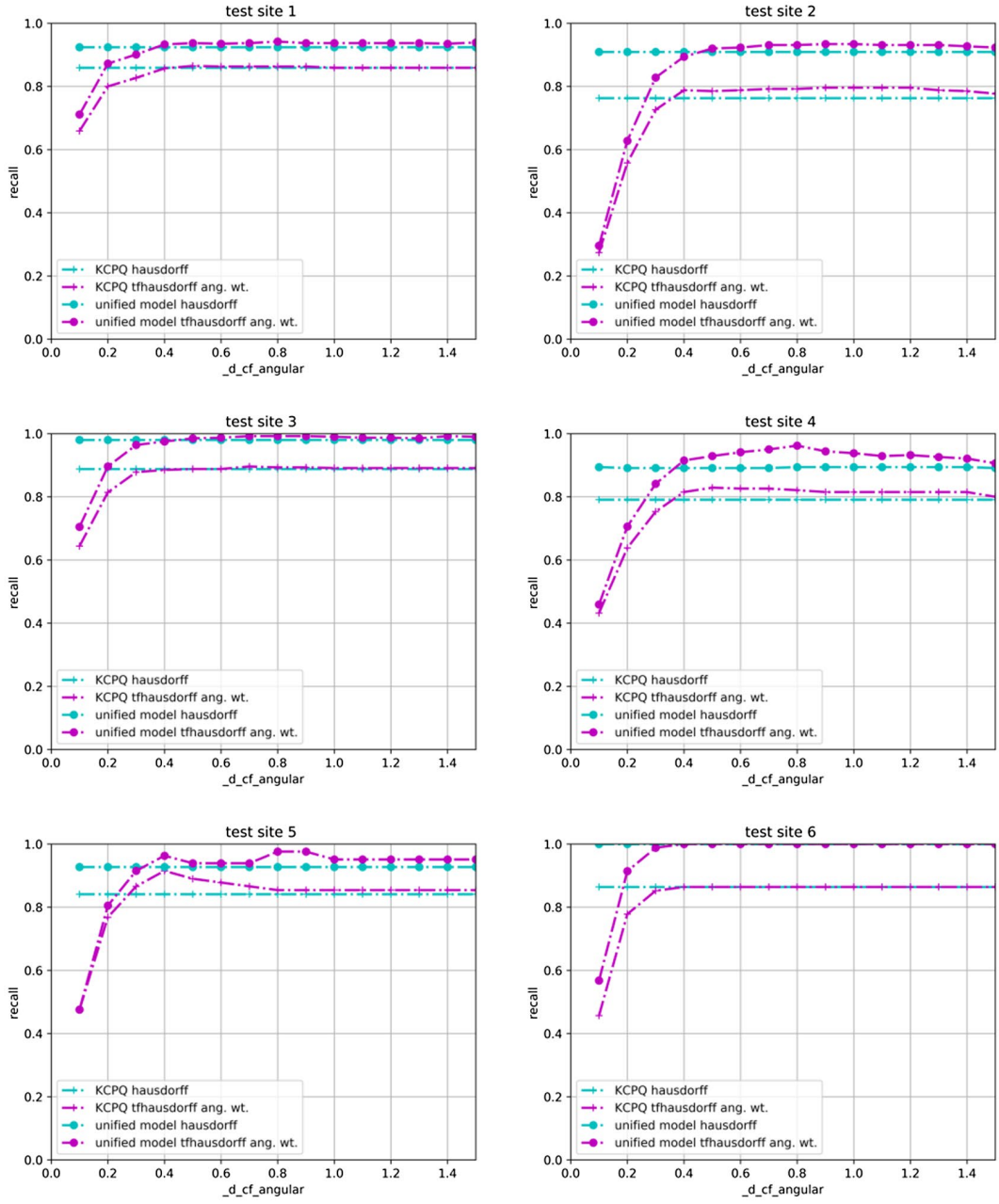

FIGURE 8 Recall rates for the KCPQ and unified conflation models on six test sites, with angular weight 0 and angular cutoff ranging over 0.1, 0.2, ..., 1.5 radians

require many parameters, except for cutoff distances to narrow down the search space. The composite DTFH distance requires additional parameters, including angular weight values. Next, we test the performance of the new distance metric versus the plain Hausdorff distance.

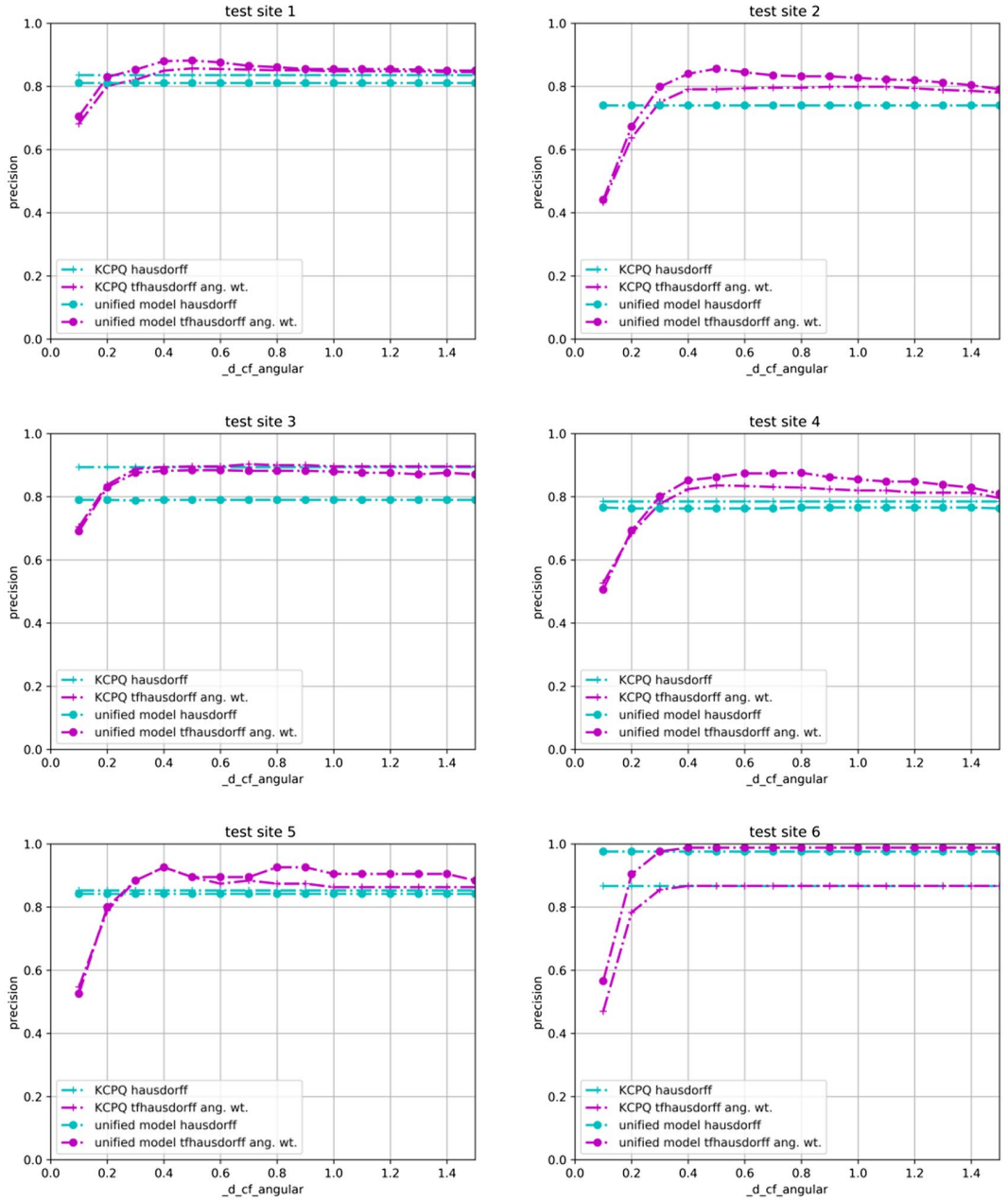

FIGURE 9 Precision for the KCPQ and unified conflation models on six test sites, with an angular weight of 0 and angular cutoff ranging over 0.1, 0.2, ..., 1.5 radians

4.2 | Performance of the composite distance metric

To evaluate the effectiveness of the composite DTFH distance, we computed the recall and precision rates for the above-mentioned conflation models under different parameters. Figure 8 presents the test results for recall rates of the tested models versus angular cutoff values, for each test site in Figure 7. We used the standard directed Hausdorff distance as the base case similarity measure for all tested models. For the proposed DTFH distance, we

initially kept the angular weight at 0 as a baseline and tested a series of angular cutoff distances in 0.1, 0.2, ..., 1.5 radians. We stop at 1.5 radians (approximately $\pi/2$) because it is the maximum possible rotation and also a large value for angular shape difference. The offset cutoff and total cutoff values are set at 100 and 150 m, respectively, as these are what we found to be sufficiently large values.

Not surprisingly, the recall rates (Figure 8) for models with the composite TF Hausdorff distance dropped compared to those of the original Hausdorff distance, since angular cutoff distances were used to remove candidate matches. This is especially true for stringent angular cutoff values of 0.3 radians or below, for which recall rates of the composite DTFH distance dropped below those of the baseline for all six test sites. For angular cutoff values of

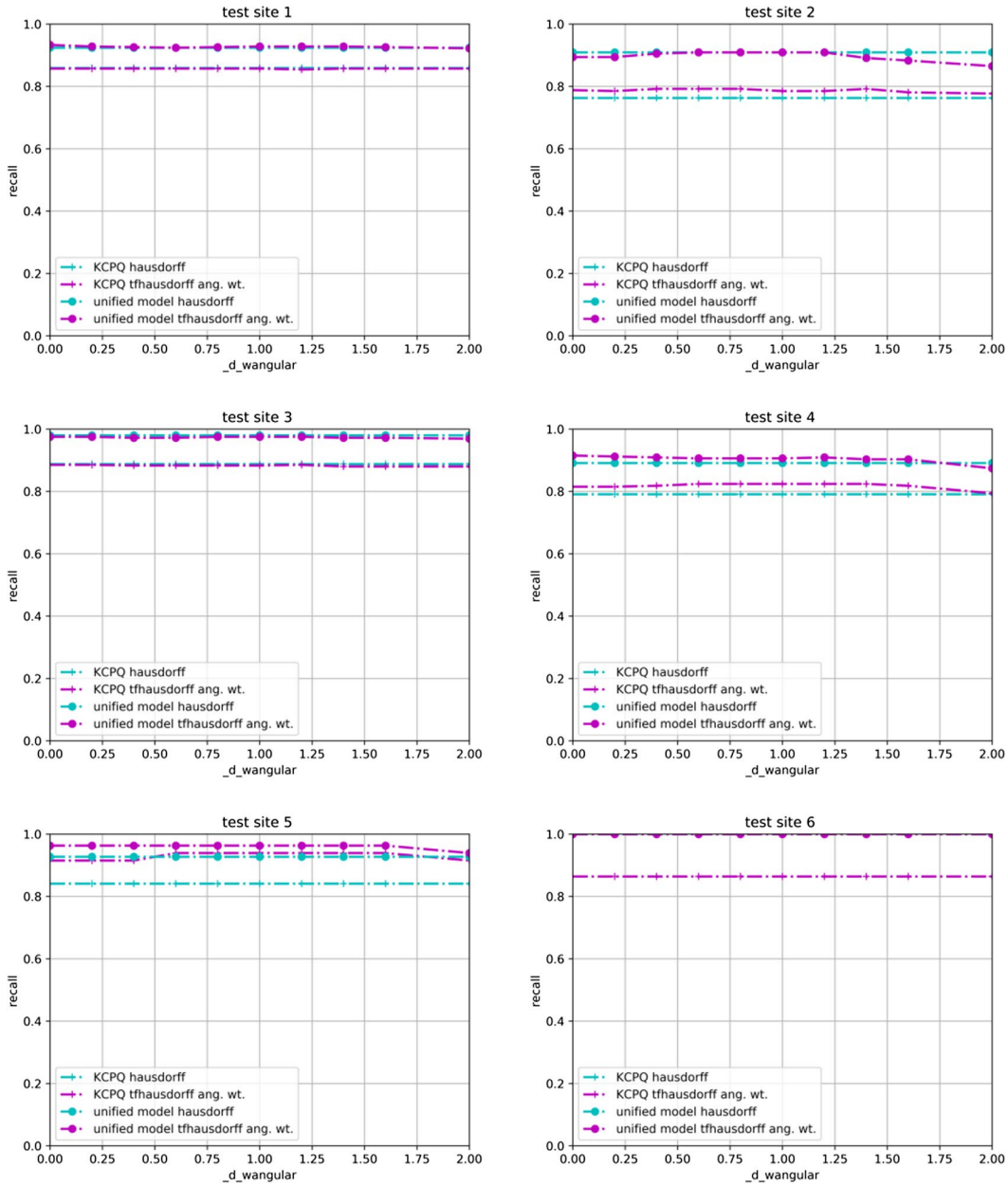

FIGURE 10 Recall rates for the KCPQ and unified conflation models with angular weight from 0 to 1.0

0.4 radians or greater, we observe no drop in recall rates. All that was captured by the standard Hausdorff distance is captured by the composite DTFH distance.

Somewhat surprisingly though, the recall rates for the composite distance are actually higher than those of the standard Hausdorff distance for several sites with suitable angular cutoffs. For sites 4 and 5, at angular cutoff 0.4, recall increased by 2.4 and 7.4% respectively for the KCPQ model, and by 2.6 and 3.6% respectively for the unified model. This could be explained by the fact that in removing unlikely matches, the composite metric can free some road features from being incorrectly assigned and thereby allow them to be assigned correctly (thereby increasing recall).

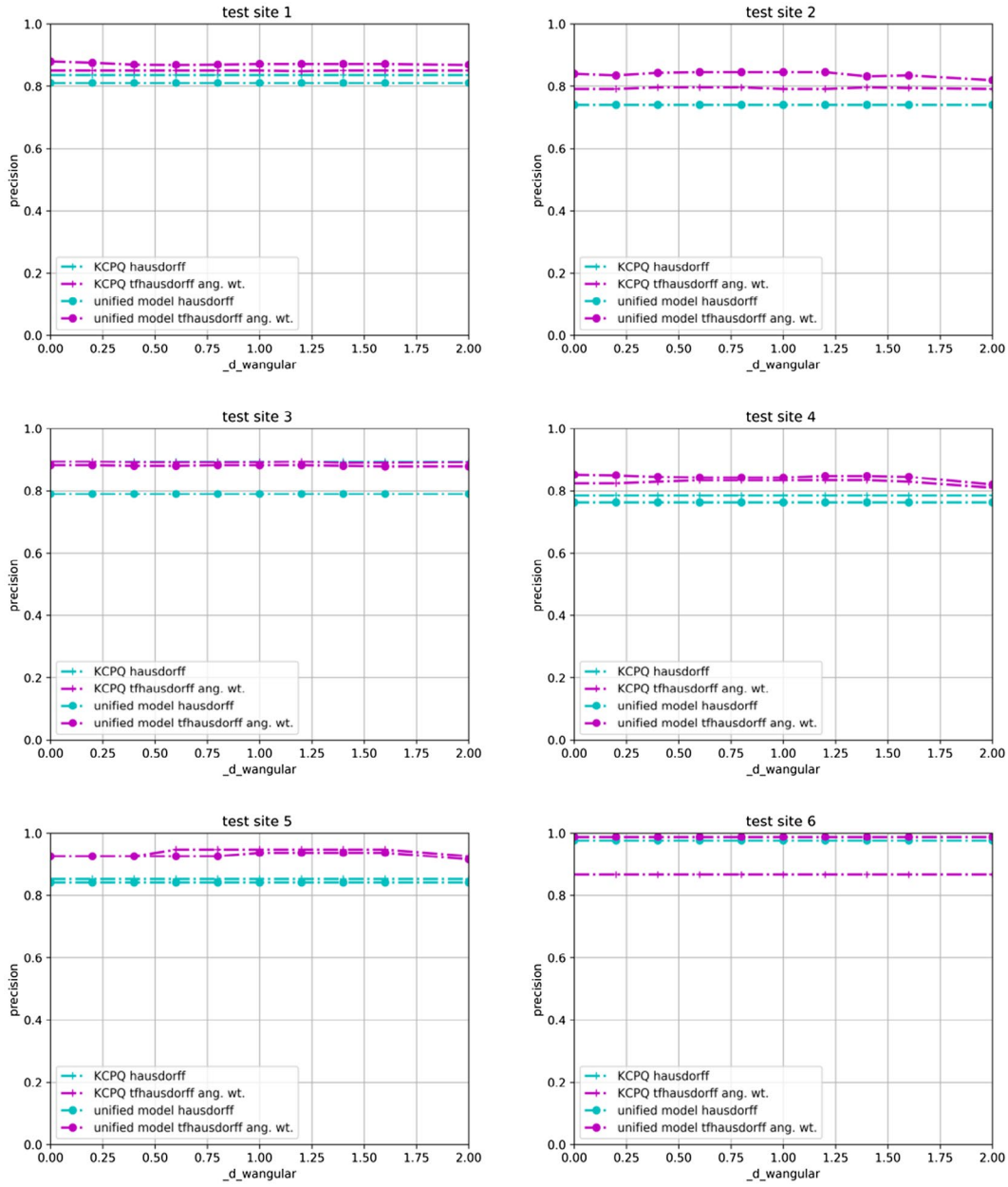

FIGURE 11 Precision rates for the KCPQ and unified conflation models with angular weight from 0 to 1.0

Comparing results between the two models, we can observe that the many-to-one unified conflation model has consistently higher recall rates than the one-to-one KCPQ model. This is because the KCPQ model is limited by its basic one-to-one assumption and cannot capture any partial matches by design. Although both models gain from using the composite metric, the many-to-one unified conflation model seems to have higher improvement in recall rates on average.

Figure 9 presents the precision rates of tested models under the same settings as Figure 8. Similar to the recall test, the precision rates for the composite distance are lower for very small angular cutoff values below 0.3 radians. However, unlike the recall rates, we can observe that with few exceptions, the precision rates of models with the composite DTFH distance are generally better than those of the models with standard Hausdorff distance when the angular cutoff is greater than 0.3 radians. For example, when the angular cutoff is 0.4 radians, the precision for the unified model has increased by 6.9, 10, 9.2, 8.9, 8.4, and 1.2% respectively for test sites 1 through 6 by using the new DTFH distance. This amounts to an average increase in precision of 7.4%. In terms of *F* score, the performance of the unified model has increased by 4.2, 5, 5.1, 6, 6.2, and 6.1% respectively for sites 1 through 6, which amounts to an average increase of 5.43% in *F* score.

The comparison of precision rates between the two conflation models is different from the situation of recall rates. For the standard Hausdorff distance, four out of six sites (sites 1, 2, 4, and 5) have similar precision rates for the unified (*m*: 1) and KCPQ (1: 1) models. This means that the significantly higher recall rates of the *m*: 1 model for these sites in Figure 8 came at a cost. Although the *m*: 1 model captures more true matches, it also introduces a greater number of false matches, which cancels out its potential contribution to precision. The difference also highlights the importance of using a strong similarity measure in the more "noisy" many-to-one conflation model.

Next, we test the sensitivity of the angular weight value w_a. Based on the previous test, we chose an angular cutoff value of 0.6 radians. Figure 10 presents the recall rates for the two tested models for a series of angular weights in 0, 0.1, … 1.0. Since the composite DTFH distance is the sum of the point-offset distance and the

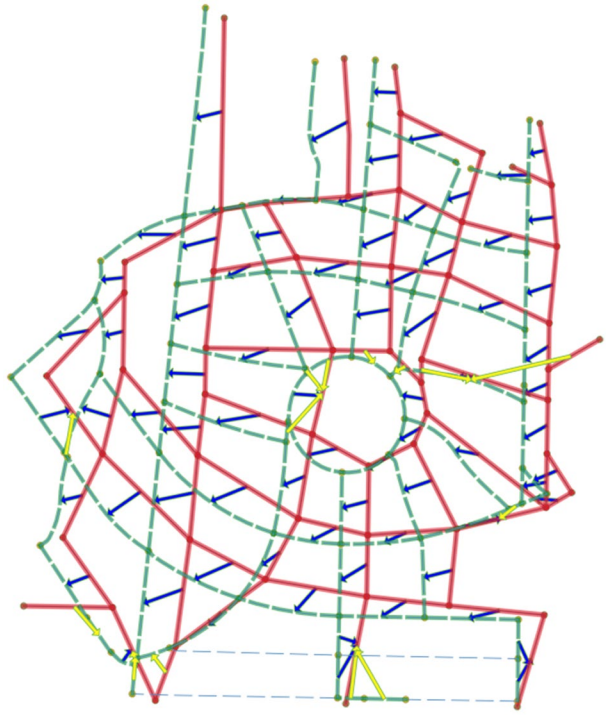

FIGURE 12 A solution of the unified conflation model with directed Hausdorff distance, for test site #5. Cutoff is 150 m. Blue/yellow arrows represent correct/erroneous matches

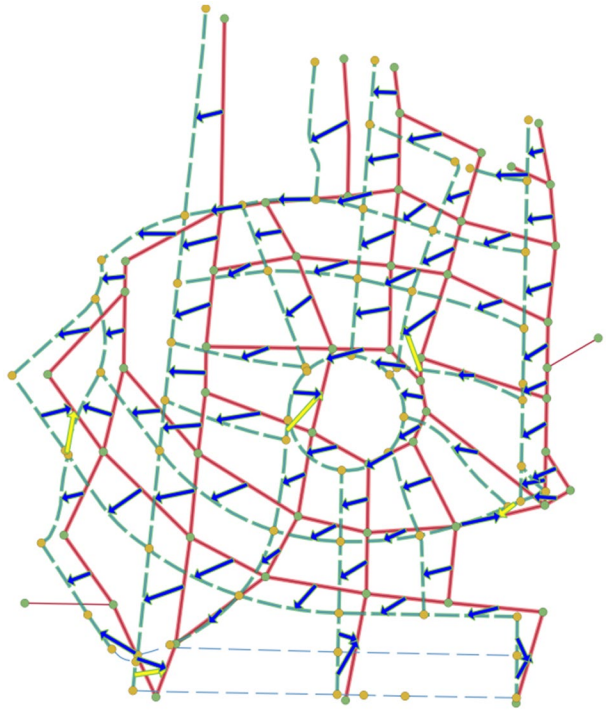

FIGURE 13 A solution of the unified conflation model with composite distance. Cutoff distance is 150 m, angular cutoff is 0.4, offset cutoff is 100 m, angular weight is 0. Blue/yellow arrows are correct/erroneous matches, respectively

weighted angular distance, a very large angular weight value may cause the composite distance to exceed the total cutoff value and therefore serve as a penalty. From Figure 10, there is no discernible drop in recall rates with the increase of angular weight values (except for a slight drop in recall for the unified model for site #2 at large angular weight values).

Figure 11 presents the precision rates for the same test (on angular weights) as Figure 10. We observe slight increases of precision rates with increasing angular weights for sites 4 and 5. Overall, using angular cutoff (c_a) values seems to be a more effective strategy in enhancing traditional distances with orientation and shape differences.

4.3 | Case studies

To illustrate the effectiveness of the new DTFH distance metric, we present in Figures 12 and 13 solutions of the unified conflation model using the standard Hausdorff distance and the new distance, respectively. The blue arrows in both figures represent correct matches and the yellow arrows represent false matches. In Figure 12, we can observe typical errors associated with the lack of angular information in the standard Hausdorff distance. Near the bottom of the map, two short horizontal lines were incorrectly matched to a street that is almost perpendicular to them. These should either have been forbidden or allowed only with great penalty. On the right edge and the lower left corner of the figure, we can observe two similar errors. By comparison, the model solution with composite TF Hausdorff distance (Figure 13) avoided most of these problems. While there are still some remaining error matches in Figure 13, they are primarily false matches associated with small segments in the OSM (green) dataset that do not have counterparts in the TIGER (red) dataset. Overall, the angular measure improved the accuracy of the match considerably.

5 | CONCLUSIONS AND FUTURE DIRECTIONS

Combining geospatial data from different sources is often required in the analysis of many spatial data. This process is known as conflation, and it is often a difficult task due to the difference in representation and level of detail in different data sources. A key factor in matching linear features such as roads and rivers is their shape and orientation. This article extends the classic TF distance into a new composite similarity metric for matching linear features in maps and GIS. The new distance metric has several advantages over the standard TF distance.

First, the new distance allows matching a linear feature to a larger counterpart feature without requiring scaling the input features. This avoids falsely matching map features that are of different sizes.

Second, the new TF distance allows the accurate location of the portion of the to-feature that corresponds to the from-feature by considering both shape deformation and orientation difference (rotation). This generalization of the location criterion in the TF distance metric allows the correct rotation in partial matches to be identified. As with the standard TF distance, we proved that the MTFD also has a finite set of positions and one only needs to evaluate these positions to compute the correct deformation and rotation.

To verify the effectiveness of the new distance metric, we compared its performance with the standard point-offset-based Hausdorff distance using two prototypical methods for handling one-to-one and one-to-many matching, respectively. Using data similar to prior work in the literature, we found that the proposed composite distance metric consistently improves the accuracy of both tested models. In particular, the precision under appropriate parameters has been improved by 7.4% on average for the one-to-many model over the six test sites.

Overall, this article demonstrates the feasibility of applying angular distance metrics for matching polyline features such as roads in GIS. Several areas of research are worthy of future investigation. Given the generality of the new distance metric, it could be applied to other types of map conflation problems such as the conflation of river networks. In addition, future work will be needed to examine the choice of the parameters, including the weight values and cutoff values for datasets in large-scale analysis. This is left as future work. While this research focuses on measuring the similarity at the element level (e.g., between individual streets), topological relationships between polylines could provide valuable information about the similarity between groups of elements (e.g., at the path level).

ORCID

Ting L. Lei https://orcid.org/0000-0003-2385-9128
Rongrong Wang https://orcid.org/0000-0002-5084-977X

REFERENCES

Ahmadi, E., & Nascimento, M. A. (2016). K-closest pairs queries in road networks. In *Proceedings of the 17th IEEE International Conference on Mobile Data Management*, Porto, Portugal (pp. 232–241). Piscataway, NJ: IEEE.

Alt, H., & Godau, M. (1995). Computing the Fréchet distance between two polygonal curves. *International Journal of Computational Geometry & Applications*, *5*(1–2), 75–91. https://doi.org/10.1142/S0218195995000064

Arkin, E. M., Chew, L. P., Huttenlocher, D. P., Kedem, K., & Mitchell, J. S. (1991). An efficiently computable metric for comparing polygonal shapes. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, *13*(3), 209–216. https://doi.org/10.1109/34.75509

Beeri, C., Kanza, Y., Safra, E., & Sagiv, Y. (2004). Object fusion in geographic information systems. In *Proceedings of the 13th International Conference on Very Large Data Bases*, Toronto, Canada (pp. 816–827). New York, NY: ACM.

Chambers, E. W., De Verdiere, E. C., Erickson, J., Lazard, S., Lazarus, F., & Thite, S. (2010). Homotopic Fréchet distance between curves or, walking your dog in the woods in polynomial time. *Computational Geometry*, *43*(3), 295–311. https://doi.org/10.1016/j.comgeo.2009.02.008

Chehreghan, A., & Abbaspour, R. A. (2017a). A new descriptor for improving geometric-based matching of linear objects on multi-scale datasets. *GIScience & Remote Sensing*, *54*(6), 836–861. https://doi.org/10.1080/15481603.2017.1338390

Chehreghan, A., & Abbaspour, R. A. (2017b). An assessment of spatial similarity degree between polylines on multi-scale, multi-source maps. *Geocarto International*, *32*(5), 471–487. https://doi.org/10.1080/10106049.2016.1155659

Church, R. L., & Garfinkel, R. S. (1978). Locating an obnoxious facility on a network. *Transportation Science*, *12*(2), 107–118. https://doi.org/10.1287/trsc.12.2.107

Devogele, T. (2002). A new merging process for data integration based on the discrete Fréchet distance. In D. E. Richardson & P. van Oosterom (Eds.), *Advances in spatial data handling* (pp. 167–181). Berlin, Germany: Springer.

Duchêne, C., Bard, S., Barillot, X., Ruas, A., Trevisan, J., & Holzapfel, F. (2003). Quantitative and qualitative description of building orientation. In *Proceedings of the Fifth ICA Workshop on Progress in Automated Map Generalisation*, Beijing, China (pp. 1–10). Washington, DC: ICA.

Eiter, T., & Mannila, H. (1994). *Computing discrete Fréchet distance*. Vienna, Austria: Technical University of Vienna.

Frank, R., & Ester, M. (2006). A quantitative similarity measure for maps. In A. Riedl, W. Kainz, & G. A. Elmes (Eds.), *Progress in spatial data handling* (pp. 435–450). Berlin, Germany: Springer.

Goodchild, M. F., & Hunter, G. J. (1997). A simple positional accuracy measure for linear features. *International Journal of Geographical Information Science*, *11*(3), 299–306. https://doi.org/10.1080/136588197242419

Lei, T. L., & Church, R. L. (2015). On the finite optimality set of the vector assignment p-median problem. *Geographical Analysis*, *47*(2), 134–145.

Lei, Z., Church, R. L., & Lei, T. L. (In review). A unified optimization model for linear feature matching in geographical data conflation. *International Journal of Geographical Information Science*.

Li, J., Li, X., & Xie, T. (2017). Morphing of building footprints using a turning angle function. *ISPRS International Journal of Geo-Information*, *6*(6), 173. https://doi.org/10.3390/ijgi6060173

Li, L., & Goodchild, M. F. (2010). Optimized feature matching in conflation. In *Proceedings of the Sixth International Conference on Geographic Information Science*, Zurich, Switzerland. http://www.giscience2010.org/pdfs/paper_111.pdf

Li, L., & Goodchild, M. F. (2011). An optimisation model for linear feature matching in geographical data conflation. *International Journal of Image & Data Fusion*, *2*(4), 309–328. https://doi.org/10.1080/19479832.2011.577458

Lombard, K., & Church, R. L. (1993). The gateway shortest path problem: Generating alternative routes for a corridor location problem. *Geographical Systems*, *1*(1), 25–45.

MacEachren, A. M. (1985). Compactness of geographic shape: Comparison and evaluation of measures. *Geografiska Annaler: Series B, Human Geography*, *67*(1), 53–67. https://doi.org/10.1080/04353684.1985.11879515

Mascret, A., Devogele, T., Le Berre, I., & Hénaff, A. (2006). Coastline matching process based on the discrete Fréchet distance. In A. Riedl, W. Kainz, & G. A. Elmes (Eds.), *Progress in spatial data handling* (pp. 383–400). Berlin, Germany: Springer.

Masuyama, A. (2006). Methods for detecting apparent differences between spatial tessellations at different time points. *International Journal of Geographical Information Science*, *20*(6), 633–648. https://doi.org/10.1080/13658810600661300

Matisziw, T. C., & Demir, E. (2016). Measuring spatial correspondence among network paths. *Geographical Analysis*, *48*(1), 3–17. https://doi.org/10.1111/gean.12078

McKenzie, G., Janowicz, K., & Adams, B. (2014). A weighted multi-attribute method for matching user-generated Points of Interest. *Cartography & Geographic Information Science*, *41*(2), 125–137. https://doi.org/10.1080/15230406.2014.880327

Ruiz, J. J., Ariza, F. J., Ureña, M. A., & Blázquez, E. B. (2011). Digital map conflation: A review of the process and a proposal for classification. *International Journal of Geographical Information Science*, *25*(9), 1439–1466. https://doi.org/10.1080/13658816.2010.519707

Ruiz-Lendínez, J. J., Ariza-López, F. J., & Ureña-Cámara, M. A. (2013). Automatic positional accuracy assessment of geospatial databases using line-based methods. *Survey Review*, *45*, 332–342. https://doi.org/10.1179/1752270613Y.0000000044

Tang, W., Hao, Y., Zhao, W., & Li, N. (2008). Research on areal feature matching algorithm based on spatial similarity. In *Proceedings of the 2008 Chinese Control and Decision Conference*, Yantai, China (pp. 3326–3330). Piscataway, NJ: IEEE.

Tong, X., Liang, D., & 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. https://doi.org/10.1080/13658816.2013.876501

Tong, X., Shi, W., & Deng, S. (2009). A probability-based multi-measure feature matching method in map conflation. *International Journal of Remote Sensing*, *30*(20), 5453–5472. https://doi.org/10.1080/01431160903130986

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

Wentz, E. A. (1997). Shape analysis in GIS. In *Proceedings of Auto-Carto 13*, Seattle, WA (pp. 7–10). Albuquerque, NM: CAGIS.

Xavier, E. M., Ariza-López, F. J., & Ureña-Cámara, M. A. (2016). A survey of measures and methods for matching geospatial vector datasets. *ACM Computing Surveys*, *49*(2), 39:1–39:34.

Xavier, E. M., Ariza-López, F. J., & Ureña-Cámara, M. A. (2017). MatchingLand, a geospatial data testbed for the assessment of matching methods. *Scientific Data*, *4*, 170180.

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

Zhang, M. (2009). *Methods and implementations of road-network matching* (Unpublished PhD dissertation). Technical University of Munich, Munich, Germany.

Zhang, X., Zhao, X., Molenaar, M., Stoter, J., Kraak, M.-J., & Ai, T. (2012). Pattern classification approaches to matching building polygons at multiple scales. *ISPRS Annals of Photogrammetry, Remote Sensing & Spatial Information Sciences*, *22*(1–2), 19–24. https://doi.org/10.5194/isprsannals-I-2-19-2012

How to cite this article: Lei T, Wang R. Conflating linear features using turning function distance: A new orientation-sensitive similarity measure. *Transactions in GIS*. 2021;25:1249–1276. https://doi.org/10.1111/tgis.12726