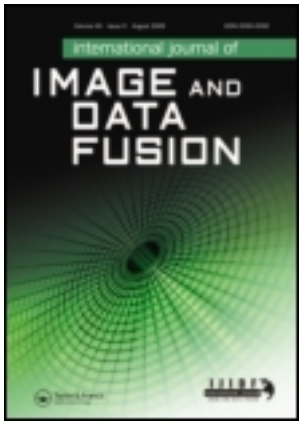



International Journal of Image and Data Fusion

Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/tidf20

# An optimisation model for linear feature matching in geographical data conflation

Linna Li [a] & Michael F. Goodchild [a]

[a] Center for Spatial Studies and Department of Geography, University of California, Santa Barbara, USA

Available online: 05 Jul 2011

To cite this article: Linna Li & Michael F. Goodchild (2011): An optimisation model for linear feature matching in geographical data conflation, International Journal of Image and Data Fusion, 2:4, 309-328

To link to this article: http://dx.doi.org/10.1080/19479832.2011.577458

*International Journal of Image and Data Fusion*
Vol. 2, No. 4, December 2011, 309–328

Taylor & Francis
Taylor & Francis Group

# An optimisation model for linear feature matching in geographical data conflation

Linna Li* and Michael F. Goodchild

*Center for Spatial Studies and Department of Geography, University of California, Santa Barbara, USA*

(*Received 3 January 2011; final version received 24 February 2011*)

Issues of heterogeneity and incompatibility in geospatial data become increasingly important as data sources become more abundant. Scientific research and decision-making usually require geospatial data from a variety of sources, since it is not realistic to collect all data directly; therefore, it is important to effectively utilise data created by various agencies using different methodologies under different circumstances. The term conflation here refers to the problem of combining incompatible geospatial data. One crucial component in conflation is feature matching, which is a prerequisite for the subsequent steps such as feature transformation. Although previous research has provided different methods of feature matching for specific applications, most of them have relied on a greedy strategy to execute the matching process. This article develops a new optimisation model to improve linear feature matching in situations with one-to-one, one-to-many and one-to-none correspondences by extending the optimised feature matching method proposed by Li and Goodchild (Automatically and accurately matching objects in geospatial datasets. *In*: *Proceedings of theory, data handling and modelling in geospatial information science*. Hong Kong, 26–28 May, 2010). Considering all possible matched pairs simultaneously, this new model achieves a high percentage of correctly matched features by maximising the total similarity between all matched pairs. When autocorrelated distortions exist in the datasets, an affine transformation can be integrated into the feature matching to improve the matching results. In addition, this study takes advantage of the asymmetry of a dissimilarity metric – directed Hausdorff distance – to address one-to-many correspondences.

**Keywords:** feature matching; conflation; optimisation; directed Hausdorff distance; affine transformation



## 1. Introduction

The rapid development of remote sensing and other technologies, as well as the growth of the Internet, make it possible to collect and access vast volumes of geographical data. Examples of well-known datasets provided by government agencies include US Census TIGER/Line files and USGS topographic maps, while companies such as TeleAtlas, Navteq, Geo-Eye and DigitalGlobe offer data services in the private sector. Meanwhile, many applications of GIS require data from more than one source; these data may be initially created for a variety of purposes, in different formats, and at various scales.

*Corresponding author. Email: linna@geog.ucsb.edu

ISSN 1947–9832 print/ISSN 1947–9824 online

http://dx.doi.org/10.1080/19479832.2011.577458
http://www.tandfonline.com

For example, in a disaster of the magnitude of Hurricane Katrina, we see that effective disaster management requires coordination of a wide range of geographical data such as that obtained from digital elevation models (DEMs), land use patterns, identification of facility locations and so forth. Transportation planning, as another example, involves the use of traffic assignment and other models that require a large amount of data input from multiple agencies. In these activities, effective integration of data in different formats is crucial for decision-making. Users of these data products may need geospatial data and other related data to be displayed as an integrated dataset for knowledge discovery. Instead of simply being able to display all the related data in the same workspace using a visual overlay operation, we must often combine these data to provide additional inferred information under a consistent and integrated data framework. In other applications, update of an existing database is required by adding new data or modifying some outdated information in the original dataset.

Conflation is the process of combining information from two or more related datasets, and thus acquiring knowledge that cannot be obtained from any single data source alone. The difficulty of this process depends on many factors, such as the complexity of representation and the size and accuracy of the involved datasets. Specifically, incompleteness and inaccuracy of the original datasets, different reference systems, distinct generalisations and representations of reality, varied scales and different purposes, as well as various time frames all create challenges in the conflation process. For example, a river may be represented as a polygon in one map, while the same river is described as a polyline in another dataset with a coarser scale and a different thematic focus.

There are generally two components in conflation: feature matching and feature transformation. Feature matching involves the identification of features in multiple datasets that represent the same entity in reality. Due to different data quality, different scales, different schemata and different purposes, the same entity may be represented differently in terms of position, shape and level of detail. This article focuses on the first component of conflation: feature matching. We use a similarity measurement that combines a directed Hausdorff distance, angle and name dissimilarity between features and achieves a satisfactory matching percentage in case studies using the optimisation model that we propose. However, other matching criteria, such as the feature type, may also be used to complement the similarity measure if available.

In geographic data conflation, it is inevitable that there are positional discrepancies between two datasets because of inherent spatial uncertainty. Kiiveri (1997) developed a model of positional uncertainty to describe positional distortion for datasets with random errors. Funk *et al.* (1999) proposed the concept of distortion field to characterise the pattern of distortion between two geographic datasets. Two kinds of positional distortion may exist in geographic datasets: random distortion and systematic distortion. Random distortion occurs when the positions of features in one dataset are randomly and independently shifted from their positions in the other dataset, due perhaps to errors of measurement. Systematic distortion occurs when the shifts in position are correlated, perhaps because many features were located through photogrammetry and inherited the same errors of image misregistration. This type of distortion can be modelled by a transformation function.

Since a rigorous consistent distortion rarely exists in two datasets, we discuss two kinds of positional distortion in this article: independent distortion and autocorrelated distortion. Autocorrelated distortion is defined as positional distortion that is strongly

spatially autocorrelated within a bounded area where Tobler's First Law of Geography is valid (Tobler 1970). Although it may not be perfectly represented by a distortion model, it can be approximated by a set of transformation functions. According to the type of distortion, two feature matching models are presented in Sections 4 and 5.

In the remainder of this article, we first provide an overview of existing methods for feature matching in Section 2. In Section 3 we present a similarity measurement for linear features that consists of a combination of geometrical and semantic information since the focus of this article is linear features, not areal features. The reason for choosing linear features as an example is that this type of data is widely used in geographical applications such as traffic assignment and navigation, and network data are one of the most difficult cases to deal with. While the feature matching model may be modified to match areal features, the measurement for calculating the similarity between areal features would be different. In Section 4 we describe our optimisation model for feature matching in linear datasets with independent distortion, given a selected similarity measurement. Section 5 incorporates affine transformation into the feature matching model to improve the matching results when autocorrelated distortion is present. Section 6 demonstrates the application of our model to six test areas of street networks and reports the matching results. Finally, we draw some conclusions based on these experiments and suggest possible future research directions in Section 7.



## 2. Background

There are generally two major steps in feature matching: First, we choose a similarity measurement to be used as a criterion for matching; second, we identify all matched pairs of features using this selected similarity criterion. If two features in different databases are represented similarly in terms of positions, shapes and relationships with surrounding features, it is probable that they represent the same entity in the real world and that the small difference between them is caused by different data schemata or uncertainty introduced in the data creation process. Samal *et al.* (2004) summarised a set of possible similarity measures that have been developed in a variety of disciplines and might be useful in conflation, including categorical similarity, string similarity and shape similarity. More broadly, similarity measures commonly used in feature matching can be classified into three types according to whether they are based on similarities of geometry, attribute or topology, or combinations of these.

Most literature in GIScience aims to address geometric feature matching, and several metrics have been developed based on distances between features, the relative directions of features and similarities of shape (Harvey 1994, Lemarié and Raynal 1996, Bel Hadj Ali 1997, Harvey and Vauglin 1997, Vauglin and Bel Hadj Ali 1998, Devogele 2002). The most widely used criteria are those based on the distance between features, and whether or not features are nearest neighbours. Feature distance is usually determined by the Euclidean distance for points or Hausdorff distance for polylines (Abbas 1994, Yuan and Tao 1999). The Euclidean distance between two points $p_1$ $(x_1, y_1)$ and $p_2$ $(x_2, y_2)$ is defined as

$$d(p_1, p_2) = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2} \tag{1}$$

The Hausdorff distance is commonly used to calculate the separation between linear features that are composed of connected straight line segments. If $d(x, L) = \min\{d(x, y) : y \in L\}$ is the shortest distance between a point $x$ and a polyline $L$, the Hausdorff distance between two such polylines, $L_1$ and $L_2$, can be defined as

$$d_H(L_1, L_2) = \max\{d_{12}, d_{21}\} \quad (2)$$

where $d_{ij} = \max_{x \in L_i}\{d(x, L_j)\}$ is the distance from $L_i$ to $L_j$. Vauglin and Bel Hadj Ali (1998) extended a Hausdorff distance model through a set of statistics to determine feature matching.

In addition to Euclidean and Hausdorff distances, other distance definitions have also been used. Devogele (2002) used the maximal distance between two lines, called the discrete Frechet distance, to consider the locations and ordering of the constituent points of polylines when calculating proximity. Bel Hadj Ali (1997) proposed an areal distance as one minus the ratio of the intersection to the union of two areas to calculate the distance between two polygons. Harvey (1994) used a buffer to determine the corresponding nodes of two features in his alignment overlay algorithm and later presented a clustering algorithm for geometric matching based on a nearness criterion (Harvey and Vauglin 1997). Another proximity-based criterion to search for a possible match is the nearest neighbour pairing. Two points *i* and *j* are nearest neighbour pairs if *j* is the nearest point in Dataset 2 to point *i* in Dataset 1, and vice versa (Saalfeld 1988, Beeri *et al.* 2004). This criterion can be extended to both polylines and polygons: When two features are each other's nearest feature in the other dataset, they are regarded as mutual nearest neighbours, no matter what kind of geometry they are. Because points are the fundamental elements that compose polylines and polygons and are the simplest features to deal with, matching between linear and areal features is usually reduced to point matching by identifying control points such as nodes and vertices (e.g. Cobb *et al.* 1998, Filin and Doytsher 2000, Chen *et al.* 2006, Masuyama 2006, Safra *et al.* 2006). In addition to proximity, angular information of linear features associated with matched points may be used to filter candidate matches. For example, Quddus *et al.* (2003) utilised the angle between GPS tracks and street networks to determine the correct match from several potential matches. Walter and Fritsch (1999) used an angle difference of less than 30° as a geometric constraint to filter potential matched pairs in their street centreline matching. For areal data, proximity could be measured by the distance between area centroids or the ratio of the intersection area of two features to their union area. To take advantage of the areal intersection metric, areas could be created for point and linear features by buffering (Hastings 2008).

Attribute similarity can also be used as a criterion for feature matching. There are three kinds of attributes: feature type, feature name and other general information. Two issues are critical in attribute matching: the mapping between different classification systems, and semantic match for feature names. Comparison between feature types could be based on synonyms with the help of a thesaurus, which would solve the problem of using different terms for the same concept. For example, one dataset might use 'street' and the other might use 'avenue' to refer to the same type of feature. Several string-similarity metrics used to calculate the difference between two terms such as place names have been developed in the information retrieval field (Cohen *et al.* 2003). Hastings (2008) used geotaxonomic and geonomial metrics to compare placetype and placenames of features in conflation of digital gazetteer entries.

The third category of criteria makes use of topological information, including the connectivity of lines, and composition relationships such as the number of incident arcs at a node, the number of arcs that form a polygon and so on. Topological information is often used to narrow the search space or as a means to validate other methods. For example, when trying to find corresponding intersections in two street network datasets, Saalfeld (1988) used the number of streets connected to the intersections as a matching criterion. Filin and Doytsher (2000) developed the 'round-trip walk' approach to take into account the link between connected line segments. If we assume that corresponding arcs emanating from counterpart nodes have counterpart nodes at the other end of the arcs (called connected nodes), this topological information can be used in feature matching. Suppose that there are two datasets Dataset 1 and Dataset 2. First, they searched all connected nodes (represented as $\{b_1, b_2, \ldots, b_k\}$) for a node $b$ in Dataset 2 that is the corresponding node for the node $a$ in Dataset 1. Then they searched back from those identified connected nodes $\{b_1, b_2, \ldots b_k\}$ in Dataset 2, identifying their corresponding nodes in Dataset 1, trying to find the connected node $a$' for those corresponding nodes in Dataset 1. If node $a$ is equal to node $a$', then these two nodes $a$ and $b$ are matched nodes.

Although the criteria for feature matching vary in different applications, a common strategy in most work is the sequential workflow of matching, called the greedy method: pairs of matched features are identified one after another. An obvious issue with this kind of method is that once a feature is matched to a wrong feature in the other dataset, no remedy can be made to correct this error. As demonstrated on two real street datasets by Li and Goodchild (2010), the percentage of correct matches using optimised feature matching is 8–9% higher than using the greedy method given the same criterion. While the greedy method intends to achieve local optimum at each step of matching, the optimisation method aims to maximise correct matches globally in the entire datasets.

This article is an extension of their previous work, but it is different from their method in several aspects. First, this model can be used to address a wider range of situations: we propose two approaches for addressing datasets with independent distortion and with autocorrelated distortion, respectively, which makes it applicable to a broader context. Second, the mathematical formulation for feature matching is not a simple assignment problem for addressing only one-to-one correspondences. In this model, we consider non-1 : 1 correspondences, including $m : 1$, $1 : m$, $1 : 0$ and $0 : 1$ correspondences in two datasets using a directed similarity measurement with the aid of a length constraint. This model effectively addresses $1 : m$ and $m : 1$ correspondences in linear features without aggregation or splitting of features in the preprocessing. Third, we incorporate spatial transformation into the feature matching process for datasets with autocorrelated distortion using an affine transformation, which greatly increases the percentage of correct matches. Fourth, the similarity measurement for feature matching has been improved – we incorporate more information that provides a more accurate metric for the similarity between features, such as a directed similarity measurement defined as a combination of a directed Hausdorff distance, angle and the dissimilarity between feature names. We take advantage of the asymmetry of directed Hausdorff distance to address $m : 1$ and $1 : m$ correspondences. Finally, we use six test areas for testing the proposed model, and these study areas represent different spatial patterns in the road network data, including grid patterns in urban areas and more irregular patterns in rural areas.

## 3. Similarity measurement

In feature matching, we need to develop an objective function whose value is an effective indicator of global goodness for making matches. A natural choice would be to maximise the total similarity between corresponding features. As stated above, there are different kinds of information that could be used in the similarity measurement. In our study, a similarity index is created by a combination of geometric and semantic information: directed Hausdorff distance, angle and dissimilarity between feature names. Hausdorff distance is selected because it characterises the proximity of two linear features particularly well (Abbas 1994). Angle of linear features is incorporated as a constraint in a directed Hausdorff distance: If the angle between two linear features is larger than a certain degree, the directed Hausdorff distance is set to a very large number, so as to prevent them from incorrectly matching with each other.

There are several cases of feature matching in terms of cardinality of matching pairs: 1 : 1 correspondence, 1 : 0 and 0 : 1 correspondences, 1 : $m$ and $m$ : 1 correspondences, and $m$ : $n$ correspondence. For 1 : 1 correspondence, each entity in reality is usually represented as only one feature in the dataset; this is the most common type (*e.g.*, $a_1$ and $b_1$, $a_2$ and $b_2$, $a_3$ and $b_3$ in Figure 3). The 1 : 0 and 0 : 1 correspondences indicate that some features in one dataset are not represented in the other dataset, either because of blunders or different time frames ($a_6$ does not have a corresponding feature in the blue dataset in Figure 4). The 1 : $m$, $m$ : 1 and $m$ : $n$ correspondences usually imply that one entity with a distinct boundary in reality is represented as more than one feature in one or both datasets, and the discretisation of entity representation is arbitrary, resulting in discrepancies in the number of features. One example of $m$ : 1 correspondence is displayed in Figure 4: $a_3$, $a_4$ and $a_5$ all correspond to $b_3$. For man-made linear features such as streets, $m$ : $n$ correspondence is not very common if we only discretise features at intersections. That is to say, one street segment with the same name is represented as only one feature in a dataset if it does not intersect with another street segment.

In order to search for $m$ : 1 and 1 : $m$ correspondences, we use two similarity values between each pair of linear features. Because of the asymmetry of the Hausdorff distance, directed Hausdorff distance from feature $i$ to feature $j$ $\mathrm{d}_{i\longrightarrow j}^{DH}$could be different from that from feature $j$ to feature $i$ $\mathrm{d}_{j\longrightarrow i}^{DH}$. Consequently, two directed Hausdorff distances are defined as

$$\mathrm{d}_{i\longrightarrow j}^{DH} = \max_{x\in L_i}\{\mathrm{d}(x, L_j)\} \tag{3}$$

$$\mathrm{d}_{j\longrightarrow i}^{DH} = \max_{x\in L_j}\{\mathrm{d}(x, L_i)\} \tag{4}$$

where $\mathrm{d}(x, L) = \min\{\mathrm{d}(x, y) : y \in L\}$ is the shortest distance between a point $x$ and a polyline $L$. This asymmetry provides a very good way to identify part and whole relationship between two linear features. For example, both $a_1$ and $a_2$ are a part of $b_1$ in Figure 1. Due to the difference in length in these linear features, the directed Hausdorff distance from $a_1$ to $b_1$ and that from $a_2$ to $b_1$ are smaller than that from $b_1$ to $a_1$, as well as that from $b_1$ to $a_2$. Accordingly, the similarity from $a_1$ to $b_1$ ($s_{\mathrm{a1,b1}}$) and that from $a_2$ to $b_1$ ($s_{\mathrm{a2,b1}}$) would be larger than that from $b_1$ to $a_1$ ($s_{\mathrm{b1,a1}}$), as well as that from $b_1$ to $a_2$ ($s_{\mathrm{b1,a2}}$). In this simple example, the similarity indices $s_{\mathrm{a1,b1}}$, $s_{\mathrm{a2,b1}}$, $s_{\mathrm{b1,a1}}$ and $s_{\mathrm{b1,a2}}$ are 110.73, 117.4, 63.24 and 87.67, respectively, using our similarity measurement.

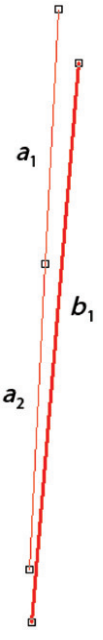



Figure 1. Directed Hausdorff distance.

The other factor in our similarity measurement is the dissimilarity between two feature names. The Hamming distance (Hamming 1950) is adopted to calculate the number of positions where corresponding alphabets are different. Because the Hamming distance is designed to compute the difference between two strings of equal length, a modified version is used in our similarity measurement to take into account feature names of different lengths in various databases. The equation for calculating the dissimilarity between feature names is defined as

$$D_{ij}^{n} = \frac{2D_{ij}^{h}}{L_i + L_j} * \alpha \tag{5}$$

where $D_{ij}^{n}$ is the name dissimilarity between two features $i$ and $j$, $D_{ij}^{h}$ is the Hamming distance, $L_i$ and $L_j$ are the lengths of two feature names and $\alpha$ is a factor to normalise the distance so it could be comparable with a directed Hausdorff distance when they are combined. In our experiment, $\alpha$ is the same as the distance threshold $a$ in Equation (6) beyond which the similarity between two features is regarded as 0. In this way, the range of both directed Hausdorff distance and the dissimilarity between two feature names is $[0, a]$.

To incorporate the threshold of directed Hausdorff distance into the matching criterion, we use similarity, instead of dissimilarity, in our feature matching model. If we use dissimilarity as the matching criterion, the dissimilarity between two features beyond the threshold of directed Hausdorff distance would be infinity and, thus, more difficult to represent in the model parameters. A similarity index between two features is computed according to a directed Hausdorff distance and the dissimilarity between feature names. First, a directed Hausdorff distance is used as a proximity constraint: if the directed

Hausdorff distance between two features is larger than a certain value $a$, the similarity between two features is set to 0. Second, if the name is not available in any of the involved features, only the directed Hausdorff distance is used to calculate the similarity. Third, if the directed Hausdorff distance is smaller than a certain value, and both names of the two features are available, a combined similarity index is calculated using these two dissimilarities. The weights of different dissimilarities are dependent on the characteristics of input datasets and we use equal weights on the two distances in our model. The following equation is used to calculate the similarity from feature $i$ to feature $j$.

$$s_{i\longrightarrow j} = \begin{cases} 0 & \text{if } \mathrm{d}^{\mathrm{DH}}_{\mathrm{i}\longrightarrow \mathrm{j}} > \mathrm{a} \\ a - \mathrm{d}^{\mathrm{DH}}_{\mathrm{i}\longrightarrow \mathrm{j}} & \text{if } \mathrm{d}^{\mathrm{DH}}_{\mathrm{i}\longrightarrow \mathrm{j}} < \mathrm{a} \text{ and } D^{\mathrm{n}}_{\mathrm{ij}} \text{ is not available} \\ a - (D^{\mathrm{n}}_{\mathrm{ij}} + \mathrm{d}^{\mathrm{DH}}_{\mathrm{i}\longrightarrow \mathrm{j}})/2 & \text{if } \mathrm{d}^{\mathrm{DH}}_{\mathrm{i}\longrightarrow \mathrm{j}} < \mathrm{a} \text{ and } D^{\mathrm{n}}_{\mathrm{ij}} \text{ is available} \end{cases} \tag{6}$$

where $s_{i\longrightarrow j}$ is the directed similarity from feature $i$ to feature $j$, $\mathrm{d}^{\mathrm{DH}}_{\mathrm{i}\longrightarrow \mathrm{j}}$is the directed Hausdorff distance from feature $i$ to feature $j$, $D^{n}_{ij}$ is the dissimilarity between two feature names and $a$ is a distance threshold beyond which two features are considered too far away to be matched. This threshold is also very helpful to reduce the search area in the matching process.

## 4. Feature matching in datasets with independent distortion

After the selection of an effective similarity measure, the next step is to find all corresponding features according to this criterion. As discussed above, most feature matching methods in the literature adopt a sequential greedy strategy, although they use a variety of criteria to decide whether two features should be matched (e.g. Cobb *et al.* 1998, Filin and Doytsher 2000, Chen *et al.* 2006). To overcome the short-sightedness of greedy methods, we identify all pairs of matched features simultaneously using an optimisation model. The goal is to find the global optimal solution from all possible choices, by maximising an objective function subject to a set of constraints. Thus, feature matching can be formulated mathematically as follows. The objective function is

$$\text{Maximise} \sum_{i=1}^{p} \sum_{j=1}^{q} s_{i\longrightarrow j} z_{i\longrightarrow j} \tag{7}$$

where $i$,$j$ are indices for the features in the first and second dataset, respectively, $p$ and $q$ are the number of features in each dataset, and $s_{i\longrightarrow j}$ is the directed similarity from feature $i$ in the first dataset to feature $j$ in the second dataset. The objective function maximises the total similarity between all matched feature pairs. The variable $z_{i\longrightarrow j}$ represents a match between feature $i$ and feature $j$, taking value 1 if a match is made and 0 otherwise, i.e.

$$z_{i\longrightarrow j} = \begin{cases} 1, & \text{if a match is made from feature } i \text{ to feature } j \\ 0, & \text{otherwise} \end{cases} \tag{8}$$

In order to find both $m:1$ and $1:m$ correspondences, we develop two sub-models using similar formulations; the only differences between these sub-models are the direction of the similarity measurement due to the asymmetry of the directed Hausdorff distance and the constraints on the objective function. Because we allow multiple features in Dataset 1 to be matched with the same feature in Dataset 2 using the directed similarity measure, $m:1$ correspondences will be identified. Similarly, if we want to identify $1:m$ correspondences,

we use the directed similarity from Dataset 2 to Dataset 1, and allow multiple features in Dataset 2 to be matched with the same feature in Dataset 1. This explains why two sub-models are necessary to complete the matching process. In Sub-Model 1, in addition to the objective function (Equations (7) and (8)), the following constraints are used:

$$\sum_{j=1}^{q} z_{i\longrightarrow j} \leq 1 \quad \forall i \tag{9}$$

$$\sum_{i=1}^{p} z_{i\longrightarrow j} + \delta_j \geq 1 \quad \forall j \tag{10}$$

The first constraint (Equation (7)) ensures that each feature in Dataset 1 must be assigned to one or no feature in Dataset 2. If the directed similarities between a feature $i_e$ in Dataset 1 and features in Dataset 2 are all very small, it is possible that this feature $i_e$ is not assigned to any feature in the other dataset, which takes care of features in Dataset 1 that do not have corresponding features in Dataset 2. The second constraint (Equation (8)) ensures that each feature in Dataset 2 must be assigned to one or more features in Dataset 1. Since more than one feature in Dataset 1 are allowed to be assigned to the same feature in Dataset 2, $m:1$ correspondence will be identified. Because it is possible that a feature $j_f$ in Dataset 2 has no corresponding features in Dataset 1, it is not reasonable to force every feature in Dataset 2 to be matched with features in Dataset 1. Therefore, we introduce a binary slack variable $\delta_j$ in the second constraint.

$$\delta_j = \begin{cases} 1, & \text{if all similarities of feature } j \text{ are less than a certain value} \\ 0, & \text{otherwise} \end{cases} \tag{11}$$

This slack variable prevents a feature in Dataset 2 from being wrongly matched with features in Dataset 1 when there is no corresponding feature. When a feature $j$ in Dataset 2 has very small similarities with all features in Dataset 1, it is acceptable that feature $j$ is not assigned to any feature in the other dataset.

Further, we use length as a constraint to prevent too many features in Dataset 1 from matching a single feature in Dataset 2. Without this constraint, each feature in Dataset 1 would be matched to its most similar feature (which is usually the closest one) in Dataset 2.

$$\sum_{i=1}^{p} l_i * z_{i\longrightarrow j} \leq k_j * \beta \quad \forall j \tag{12}$$

where $l_i$ is the length of feature $i$ in Dataset 1, $k_j$ is the length of feature $j$ in Dataset 2, and this constraint ensures that the total length of all features in Dataset 1 that are matched to the same feature $j$ in Dataset 2 does not exceed the length of feature $j$ times $\beta$. The parameter $\beta$ is a tolerance factor that takes into account uncertainty in feature length in different datasets, depending on the resolution of the two input datasets. For input datasets of similar resolution as that in our experiments, $\beta$ is a value a little larger than 1 ($\beta = 1.4$ in our model). If Dataset 2 has a much higher resolution than Dataset 1, i.e. the same linear feature has a longer length in Dataset 2, $\beta$ is smaller than 1. On the contrary, if Dataset 2 has a lower resolution than Dataset 1, $\beta$ should be assigned a value larger than 1 based on the ratio of the resolution of two datasets. However, determination of the value for $\beta$ may need a closer examination of the data for more complicated situations. In some

datasets, the length differences may not be due to the resolution difference, but rather to different cartographic generalisations. For example, in a stream network, several small streams may be simplified as one feature in one dataset, but they are represented as many features in another one.

Equations (7)–(12) describe Sub-Model 1 for feature matching that identifies 1 : 1 and $m$ : 1 correspondences from Dataset 1 to Dataset 2. Similarly, we use Sub-Model 2 to identify 1 : 1 and 1 : $m$ correspondences between the same datasets. In Sub-Model 2, the objective function and constraints are as follows where each symbol has a similar meaning to that in Sub-Model 1:

$$\text{Maximise} \sum_{i=1}^{p} \sum_{j=1}^{q} s_{j \longrightarrow i} z_{j \longrightarrow i} \tag{13}$$

$$z_{j \longrightarrow i} = \begin{cases} 1, & \text{if a match is made from feature } j \text{ to feature } i \\ 0, & \text{otherwise} \end{cases} \tag{14}$$

$$\sum_{i=1}^{p} z_{j \longrightarrow i} \leq 1 \quad \forall j \tag{15}$$

$$\sum_{j=1}^{q} z_{j \longrightarrow i} + \delta_i \geq 1 \quad \forall i \tag{16}$$

$$\delta_i = \begin{cases} 1, & \text{if all similarities of feature } i \text{ is less than a certain value} \\ 0, & \text{otherwise} \end{cases} \tag{17}$$

$$\sum_{j=1}^{q} k_j * z_{j \longrightarrow i} \leq l_i * \beta \quad \forall i \tag{18}$$

## 5. Feature matching in datasets with autocorrelated distortion

We have presented a general optimisation model for matching geospatial features in datasets with independent distortion. In this section, we present an approach for matching features in datasets with autocorrelated distortion where rectification of autocorrelated errors in original datasets leads to improvement for feature matching. The pattern of global distortion within a bounded area is taken into account by integrating an affine transformation into the objective function of our optimisation model. If we assume that the relationship between two datasets can be approximated by an affine transformation, the transformed coordinates of features in Dataset 2 can be written as

$$u'_j = a + bu_j + cv_j \tag{19}$$

$$v'_j = d + eu_j + fv_j \tag{20}$$

where $(u_j, v_j)$ are original coordinates of features in Dataset 2, $(u'_j, v'_j)$ are transformed coordinates of features in Dataset 2, and $a$, $b$, $c$, $d$, $e$, $f$ are parameters for affine transformation. Then $(u'_j, v'_j)$, rather than $(u_j, v_j)$, are used in the calculation of directed Hausdorff distance in Equations (3) (4), which will be used in the calculation of directed feature similarity and then substituted into the objective function in Equation (7).

The heuristic for solving the affine-transformation-based optimisation model adopts the following procedure.

(1) A number $k$ of pairs of non-collinear control points ($k \geq 3$) is specified.
(2) Extract start and end points of features with $\delta_i = 0$ in Dataset 1. Only features with $\delta_i = 0$ are selected because it is very likely that features with $\delta_i = 1$ have no correspondences in Dataset 2. Similarly, extract start and end points of features with $\delta_j = 0$ in Dataset 2. We reduce the search space for our model by only considering features with counterparts as control points.
(3) The study area in Dataset 1 is evenly divided into $k$ sub-regions with similar area. After the sub-regions are generated, one point is randomly selected from each sub-region as a control point ($k$ control points selected in total). This is stratified sampling to distribute control points as evenly as possible.
(4) For each control point $p_r$ in Dataset 1, find all points within a buffer of possible maximum positional offset in Dataset 2, and store these candidate control points in the set $Q_r = \{q_{r1}, q_{r2}, \ldots, q_{rs}\}$, where $q_{r1}, q_{r2}, \ldots, q_{rs}$ are all possible matched control points for $p_r$, and $s$ is the total number of possible matched points; $r = 1,2,\ldots,k$.
(5) For each control point in Dataset 1, choose a point from each of sets $Q_r$ ($r = 1,2,\ldots,k$) as a corresponding control point. For instance, in the first iteration, the first point in each set $Q_r$ is selected as the corresponding point for $p_r$. The second iteration selects the second point in $Q_r$ if there is more than one candidate in the control point set. Different iterations generate different pairs of control points. Affine transformation parameters $a$, $b$, $c$, $d$, $e$ and $f$ are calculated based on selected pairs of control points using least squares.
(6) Substitute these parameters back into the objective function, solve the optimisation model, and record the total similarity of all matched features.
(7) Repeat steps 5,6 until all possible combinations of control point pairs are exhausted. The final solution of this model obtains affine transformation parameters $a$, $b$, $c$, $d$, $e$, $f$ that generate the largest total similarity between corresponding features.

This model with different approaches for different kinds of positional distortion was implemented in GLPK (GNU Linear Programming Kit), an open-source package for solving linear programming and mixed integer programming problems.

## 6. Experiments

In this section, we test two approaches for feature matching on six test areas (Figure 2) that contain real street-network data for Santa Barbara, CA, extracted from two different sources: Topologically Integrated Geographic Encoding and Referencing system (TIGER) and Tele Atlas. The two sources have about the same resolution and level of spatial accuracy: approximately 10 m in the urban area and 50 m in the rural area. As noted in the similarity measurement, these tests emphasise the positional similarity between linear

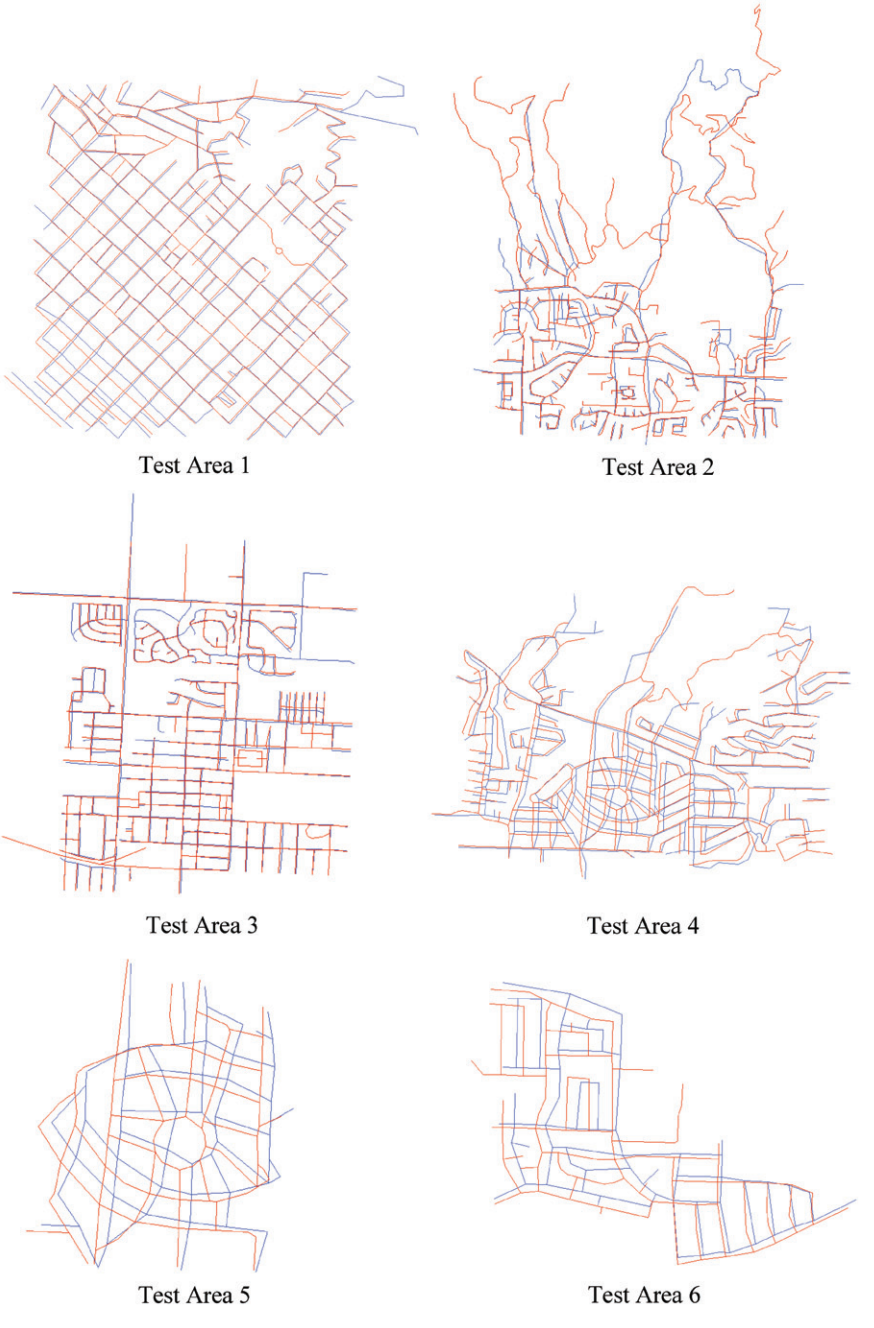



Figure 2. Test areas of street networks in Santa Barbara, CA.

features; topology, feature type and other kinds of information are not involved. These test areas represent two major types of streets in a street database: streets in both urban and rural areas. They have different street patterns: Test Area 1 represents regular grid streets in downtown Santa Barbara; Test Area 2 describes irregular streets in the hilly areas;

Test Areas 3 and 4 contain streets that have both grid patterns and curvy streets. These four areas are used to test the approach for datasets with independent distortion. Test Areas 5 and 6 contain streets that have typical autocorrelated distortion. These two test areas are smaller, because autocorrelated distortion is usually expected over a limited area in geographic datasets. Both approaches were tested on Test Areas 5 and 6. The features to be matched are street segments that are separated by street intersections. We assume that pre-processing has been done to the original datasets to ensure that individual datasets are internally consistent, and that they use the same coordinate system and projection. Six pairs of datasets have different numbers of polylines and different numbers of vertices composing polylines. As can be seen, there are apparent discrepancies between these two datasets. Some streets appear in one dataset, but not in the other one; further, the same streets are represented as different shapes at different locations.



### 6.1 *Matching results*

In evaluating matching results, we examine both corresponding feature pairs and features without correspondences. For a particular feature, there are three possibilities: (1) a feature has no corresponding features in the other dataset, and we call it *a single*; (2) a feature has one corresponding feature in the other dataset, and we call it *a perfect pair*; (3) a feature has multiple corresponding features in the other dataset, and we call it *a partial pair*. When a feature falls into the first category, and it is identified as having no correspondence, we count it as a correct identification. When a feature falls into the second category, and this pair of corresponding features is identified as being matched, it is a correct identification. When a feature falls into the third category, it forms several corresponding pairs with all of its partial counterparts, and each pair is counted as a correct identification if it is matched correctly. For example, if a feature $b_1$ in Dataset 2 has two corresponding features $a_1$ and $a_2$ in Dataset 1 (Figure 1) and $b_1$ is matched correctly to $a_1$, then the pair between $a_1$ and $b_1$ is a correct identification; and the pair between $a_2$ and $b_1$ is not a correct identification if they are not matched as a corresponding pair. Incorrect identifications are counted under the following conditions: when a feature is correctly identified as having a match, but its corresponding feature is mistakenly selected; when a feature is identified incorrectly as having a match when it actually does not (false positive), or when a feature is identified incorrectly as not having a match when it actually does (false negative). The matching results from this model were compared with the true corresponding pairs and singles identified by people. The percentage of correct identifications is used as a measure to evaluate matching performance. Table 1 demonstrates the results for the four test areas using the approach for datasets with independent distortion. The percentage of correct identifications varies from one test area to another and mostly depends on the spatial pattern of features, feature density and discrepancies between the two input datasets. The average percentage of correct identifications is 97.18%. The experiments show that the spatial pattern of the data is more important than the number of features in affecting the performance of feature matching.

Both approaches were applied in Test Areas 5 and 6 that have typical autocorrelated distortion between two datasets (Figure 2). Six pairs of control points were used in the second approach ($k = 6$) to obtain the affine transformation parameters. The results are displayed in Table 2. Test Area 5 achieves 100% correct identifications using the second

Table 1. Results of optimised feature matching for datasets with independent distortion.

| Test area | Test area 1 | Test area 2 | Test area 3 | Test area 4 | Total |
|---|---|---|---|---|---|
| Number of features in Dataset 1 | 434 | 308 | 377 | 344 | 1463 |
| Number of features in Dataset 2 | 423 | 264 | 374 | 322 | 1383 |
| Number of corresponding pairs and singles | 450 | 330 | 419 | 362 | 1561 |
| Number of correct identifications | 441 | 322 | 410 | 344 | 1517 |
| Percentage of correct identifications | 98.00 | 97.58 | 97.85 | 95.03 | 97.18 |



Table 2. Matching results for Test Areas 5 and 6: datasets with autocorrelated distortion.

| Test area | Test area 5 | Test area 6 |
|---|---|---|
| Number of features in Dataset 1 | 81 | 84 |
| Number of features in Dataset 2 | 80 | 77 |
| Number of corresponding pairs and singles | 81 | 84 |
| Number of correct identifications using Approach 1 | 74 | 79 |
| Number of correct identifications using Approach 2 | 81 | 82 |
| Percentage of correct identifications using Approach 1 | 91.36 | 94.05 |
| Percentage of correct identifications using Approach 2 | 100.00 | 97.62 |
| Computation time using Approach 1 | 0.6 s | 0.5 s |
| Computation time using Approach 2 | 0.8 s | 0.6 s |

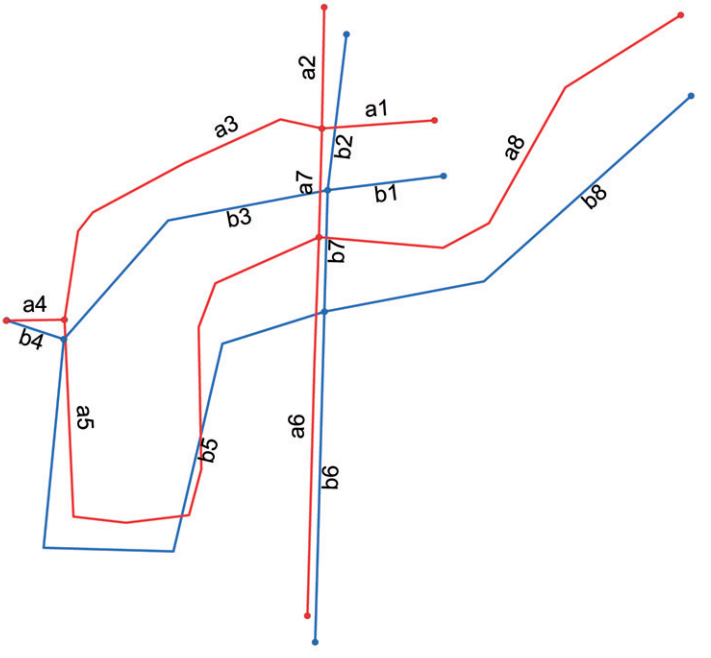
approach, a significant improvement compared to the results without consideration of autocorrelated distortion in the first approach (91.36% correct identifications). In Test Area 6, only two pairs of corresponding features are not correctly identified using Approach 2 due to the difference in feature names, which may be solved by using a more accurate metric for name dissimilarity. We use smaller test areas in these two examples, because autocorrelated discrepancies usually exist over a limited area in geographical datasets. When larger datasets are involved, we may divide the study area into several subareas using techniques like divide-and-conquer (Preparata and Shamos 1985) and solve each part simultaneously, using techniques such as piecewise distortion models in which the dataset is partitioned into several regions and the positional distortion in each region is approximated by an affine transformation (Funk *et al*. 1999).

### 6.2 *Some matching examples*

In this section some matching examples are presented to demonstrate the results of the optimisation model for featuring matching, which include 1 : 1 correspondence, $m$ : 1 correspondence, 1 : $m$ correspondence and features without correspondences.

#### 6.2.1 *1 : 1 correspondence*

If a feature in Dataset 1 has only one corresponding feature in Dataset 2 and the same feature in Dataset 2 also has only one corresponding feature in Dataset 1, this correspondence is identified in both sub-models, resulting in a 1 : 1 correspondence.




In Sub-Model 1: $a_1 -> b_1$, $a_2 -> b_2$, $a_3 -> b_3$, $a_4 -> b_4$, $a_5 -> b_5$, $a_6 -> b_6$, $a_7 -> b_7$, $a_8 -> b_8$
In Sub-Model 2: $b_1 -> a_1$, $b_2 -> a_2$, $b_3 -> a_3$, $b_4 -> a_4$, $b_5 -> a_5$, $b_6 -> a_6$, $b_7 -> a_7$, $b_8 -> a_8$

Figure 3. 1 : 1 correspondence in feature matching.

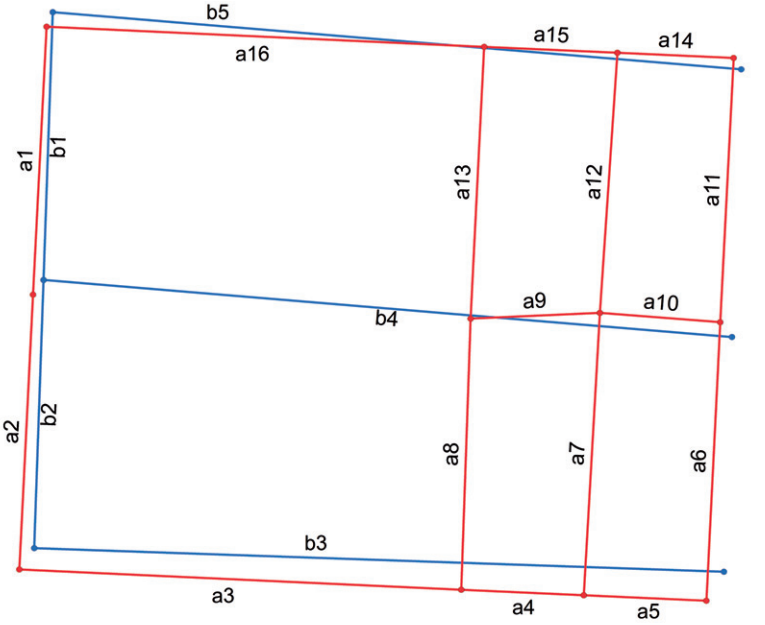
For example, there are eight pairs of 1 : 1 correspondences in Figure 3 and they are all correctly identified in both sub-models.

### 6.2.2 *m : 1 correspondence*

If multiple features in Dataset 1 have the same corresponding feature in Dataset 2, these corresponding pairs of features are identified in Sub-Model 1, but not in Sub-Model 2, due to the asymmetry of directed similarity measurement; thus $m$ : 1 correspondence is established (Figure 4).

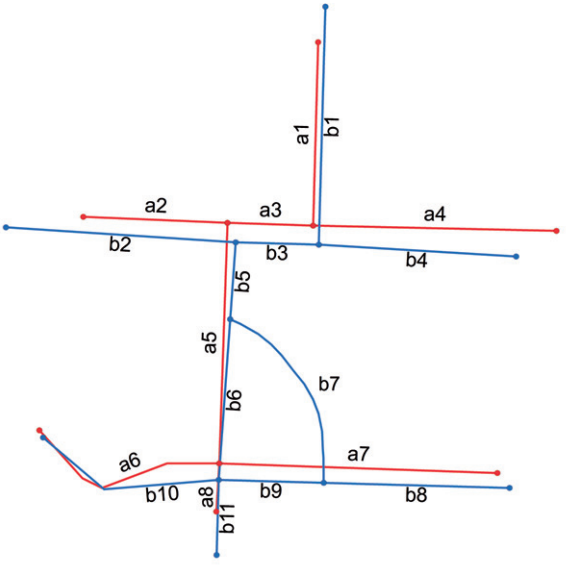
### 6.2.3 *1 : m correspondence*

If one feature in Dataset 1 has multiple corresponding features in Dataset 2, these corresponding pairs of features are identified in Sub-Model 2, but not in Sub-Model1, resulting in 1 : $m$ correspondences (Figure 5).




In Sub-Model 1: $a_1$ - > $b_1$, $a_2$ - > $b_2$, $a_3$ - > $b_3$, $a_4$ - > $b_3$, $a_5$ - > $b_3$, $a_9$ - > $b_4$, $a_{10}$ - > $b_4$, $a_{14}$ - > $b_5$, $a_{15}$ - > $b_5$, $a_{16}$ - > $b_5$
In Sub-Model 2: $b_1$ - > $a_1$, $b_2$ - > $a_2$

Figure 4. *m*:1 correspondence in feature matching.




In Sub-Model 1: $a_1$ - > $b_1$, $a_2$ - > $b_2$, $a_3$ - > $b_3$, $a_4$ - > $b_4$, $a_6$ - > $b_{10}$, $a_8$ - > $b_{11}$
In Sub-Model 2: $b_1$ - > $a_1$, $b_2$ - > $a_2$, $b_3$ - > $a_3$, $b_4$ - > $a_4$, $b_5$ - > $a_5$, $b_6$ - > $a_5$, $b_8$ - > $a_7$, $b_9$ - > $a_7$, $b_{10}$ - > $a_6$, $b_{11}$ - > $a_8$

Figure 5. 1:*m* correspondence in feature matching.

Table 3. Computation time for optimised feature matching.

| Test area | Test area 1 | Test area 2 | Test area 3 | Test area 4 |
|---|---|---|---|---|
| Computation time | 31 min 39 s | 23 min 16 s | 1 min 37 s | 2 h 6 min 14 s |

#### 6.2.4 *Features without correspondences*

If a feature in one dataset has no corresponding features in the other dataset, it is not matched to any feature; and consequently, these features are identified as having no corresponding features. For example, Features $a_6$, $a_7$, $a_8$, $a_{11}$, $a_{12}$ and $a_{13}$ are not matched to any feature in Figure 4; Feature $b_7$ is not matched to any feature in Figure 5.

### 6.3 *Computation time*

Since the main purpose of this study is not to create highly efficient programs, but to demonstrate the effectiveness of this optimisation model for feature matching, we do not focus on the improvement of the computation time, although it is possible that a more efficient algorithm may be used to solve the defined optimisation problem. Table 3 shows the computation time for our study areas. Computation time is largely dependent on the nature of the data: the spatial pattern of street networks, the number of features in the study area and the similarity between the two input datasets. We could always partition the datasets into smaller subareas using divide-and-conquer, matching features within each subarea separately. Parallel computing may also offer a potential for faster computing if a problem is divided spatially in this way.

## 7. Conclusion

Feature matching is one of the crucial components in conflation. Without correct identification of matched features in different datasets, the subsequent steps such as feature transformation will not be executed properly. To the best of our knowledge, most published feature-matching methods adopt a greedy strategy with no possibility of correcting mismatched features during execution of the algorithm. Here, we propose a new dual strategy for feature matching in conflation: The matching process is formulated as an optimisation model that takes into account all potentially matched pairs simultaneously by maximising the total similarity of all matched features; and two approaches are defined based on the nature of the positional distortions. As demonstrated in test datasets of road networks in both urban and rural areas, this strategy achieves an average of 97.18% of correct identifications. The benefit of the optimisation method is that errors typical of greedy approaches are reduced by a simultaneous matching of all feature pairs. Greedy methods usually select potential matches sequentially at each step, reducing the search space until it is empty. In contrast, our method does not need to find potential matching features in the preprocessing step and find the most similar correspondence for each feature iteratively. It is simple to implement in only one step and requires less human interaction than the greedy method. As shown in the experiments, our method achieves a good performance without using too much information: only the proximity, angle and

feature names are used. If more information is integrated, the potential to improve the percentage of correct identifications is good. Although only linear features are tested in this study, this model can be extended to areal features as well, as long as the selected similarity criteria are an adequate indicator of the resemblance between areal features. While directed Hausdorff distance is appropriate for measuring the similarity between linear features in an $m:1$ correspondence, another metric for measuring the part-and-whole similarity between areal features must be explored in future research. The success of this new model for feature matching depends on the appropriate selection of several parameters including $\alpha$ and $\beta$ according to the characteristics of input datasets. Identification of features without correspondences in input datasets provides a good opportunity to study change detection. In terms of similarity measurement, the directed Hausdorff distance has proved to be an effective metric to match a part of a linear feature to a whole feature, thus accurately identifying $1:m$ and $m:1$ correspondences without aggregation or splitting of features in the preprocessing before matching. Further, we differentiate between geographical datasets with independent and autocorrelated distortions. For datasets with autocorrelated distortions, we take into account rectification of those positional offsets by incorporating an affine transformation into the optimised feature matching model which significantly increases the percentage of correct identifications. In summary, the proposed optimised linear feature matching model achieves a high percentage of correct identifications by effectively addressing $1:1$, $1:0$, $0:1$, $1:m$, and $m:1$ correspondences with the aid of the asymmetric directed Hausdorff distance metric.

There are several research questions that merit further investigation: Foremost is the improvement of the similarity measurement, since our current method to calculate the name dissimilarity is very simple. The name dissimilarity could be improved to better represent the difference between two strings by taking into account more complex situations, such as omission of several characters in the middle of a name. For example, we may use an edit-distance, such as Levenshtein distance that counts the minimum number of edits to transform one string to another, to calculate the dissimilarity between two names (Levenshtein 1965). In addition, more information could be integrated into this similarity measurement when it is available, such as feature type and other semantic and topological information. We plan to test this model on other types of geographical data, including point datasets such as landmarks and areal data like parcels. Finally, when the input datasets are large, automatically partitioning the study area according to spatial autocorrelation and solving feature matching in each sub-area is very important for efficiency improvement and for incorporation of spatial transformation; thus, investigation of a way to divide the study area automatically according to spatial autocorrelation is important. A possible solution would be to use criteria such as a ratio of areal dependence (RAD) statistic (Funk *et al.* 1999). The RAD statistic is 1 minus the ratio of local variance at a sample point divided by the expected variance at that point in a distortion field; as a result, similar RAD values denote points of similar distortion, according to which clusters and subsequent regions of spatially autocorrelated distortion may be generated.

**Acknowledgements**

This research was supported by the US Department of Transportation (USDOT), the National Geospatial-Intelligence Agency through the NGA University Research Initiative Program (NGA-NURI grant no. HM1582-10-1-0007) and the Army Research Office (ARO).

## References


Abbas, I., 1994. *Base de données vectorielles et erreur cartographique: problèmes posés par le contrôle ponctuel; une méthode alternative fondée sur la distance de Hausdorf. Computer Science*. Paris: Université de Paris VII.

Beeri, C., *et al*., 2004. Object fusion in geographic information systems. *Proceedings of the thirtieth international conference on very large data bases* Vol. 30. August 31–September 3, Toronto, Canada, 816–827.

Bel Hadj Ali, A., 1997. *Appariement geometrique des objets géographiques et étude des indicateurs de qualité*. Saint-Mandé. Paris: Laboratoire COGIT.

Chen, C.-C., Knoblock, C., and Shahabi., C., 2006. Automatically conflating road vector data with orthoimagery. *GeoInformatica*, 10 (4), 495–530.

Cobb, M.A., *et al*., 1998. A rule-based approach for the conflation of attributed vector data. *Geoinformatica*, 2 (1), 7–35.

Cohen, W., Ravikumar, P., and Fienberg, S.E., 2003. A comparison of string distance metrics for name-matching tasks. *In*: *Proceedings of IJCAI-2003 Workshop an Information Integration on the web (IIWeb-03)*, Acapulco, Mexico, August 9–10, 2003.

Devogele, T., 2002. A new merging process for data integration based on the discrete Frechet distance. *In*: D. Richardson and P. van Oosterom, eds. *Advances in spatial data handling*. New York: Springer Verlag, 167–181.

Filin, S. and Doytsher, Y., 2000. The detection of corresponding objects in a linear-based map conflation. *Surveying and land information systems*, 60 (2), 117–128.

Funk, C., *et al*., 1999. Formulation and test of a model of positional distortion fields. *In*: K. Lowell and A. Jaton, eds. *Spatial accuracy assessment: land information uncertainty in natural resources*. Chelsea, MI: Ann Arbor Press, 131–138.

Hamming, R.W., 1950. Error detecting and error correcting codes. *Bell System Technical Journal*, 26 (2), 147–160.

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

Harvey, F., 1994. Defining unmoveable nodes/segments as part of vector overlay. *In: Sixth international symposium on spatial data handling*. Edinburgh, Scottland, T.C. Waugh, IGU Commision on GIS, Association for Geographic Information.

Harvey, F. and Vauglin, F. 1997. *No fuzzy creep! A clustering algorithm for controlling arbitrary node movement. AutoCarto* 13, Seattle, ASPRS/ASCM.

Kiiveri, H.T., 1997. Assessing, representing and transmitting positional uncertainty in maps. *International Journal of Geographical Information Science*, 11 (1), 33–52.

Lemarié, C. and Raynal, L. 1996. Geographic data matching: first investigations for a generic tool. GIS/LIS '96, Denver, Co, ASPRS/AAG/URISA/AM-FM.

Levenshtein, V., 1965. Binary codes capable of correcting spurious insertions and deletions of ones. *Problems of Information Transmission*, 1, 8–17.

Li, L. and Goodchild, M.F., 2010. Automatically and accurately matching objects in geospatial datasets. *In*: *Proceedings of theory, data handling and modelling in geospatial information science*. Hong Kong, 26–28 May, 2010.

Masuyama, A., 2006. Methods for detecting apparent differences between spatial tessellations at different time points. *International Journal of Geographical Information Science*, 20, 633–648.

Preparata, F.P. and Shamos, M.I., 1985. *Computational geometry: an introduction*. New York, NY: Springer-Verlag New York, Inc.

Quddus, M., *et al*., 2003. A general map matching algorithm for transport telematics applications. *GPS Solutions*, 7 (3), 157–167.

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

Safra, E., *et al*., 2006. Efficient integration of road maps. *In: Proceedings of the 14th annual ACM international symposium on advances in geographic information systems*, Arlington, VA, 59–66.

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.

Tobler, W., 1970. A computer movie simulating urban growth in the Detroit region. *Economic Geography*, 46 (2), 234–240.

Vauglin, F. and Bel Hadj Ali, A. 1998. Geometric matching of polygonal surfaces in GISs. ASPRS Annual Meeting, Tampa, FL, ASPRS.

Walter, V. and Fritsch, D., 1999. Matching spatial data sets: a statistical approach. *International Journal of Geographical Information Science*, 13, 445–473.

Yuan, S. and Tao, C. 1999. Development of conflation components. *The Proceedings of Geoinformatics'99 Conference*, Ann Arbor, MI.