International Journal of Geo-Information

MDPI

Article

On the Theoretical Link between Optimized Geospatial Conflation Models for Linear Features

Zhen Lei [1], Zhangshun Yuan [1] and Ting L. Lei [2,*]

[1] College of Automation, Wuhan University of Technology, Wuhan 430070, China; leizhen@whut.edu.cn (Z.L.); yuan_zhangshun@whut.edu.cn (Z.Y.)
[2] Department of Geography and Atmospheric Science, University of Kansas, Lawrence, KS 66045, USA
* Correspondence: lei@ku.edu

Abstract: Geospatial data conflation involves matching and combining two maps to create a new map. It has received increased research attention in recent years due to its wide range of applications in GIS (Geographic Information System) data production and analysis. The map assignment problem (conceptualized in the 1980s) is one of the earliest conflation methods, in which GIS features from two maps are matched by minimizing their total discrepancy or distance. Recently, more flexible optimization models have been proposed. This includes conflation models based on the network flow problem and new models based on Mixed Integer Linear Programming (MILP). A natural question is: how are these models related or different, and how do they compare? In this study, an analytic review of major optimized conflation models in the literature is conducted and the structural linkages between them are identified. Moreover, a MILP model (the *base-matching* problem) and its bi-matching version are presented as a common basis. Our analysis shows that the assignment problem and all other optimized conflation models in the literature can be viewed or reformulated as variants of the base models. For network-flow based models, proof is presented that the *base-matching* problem is equivalent to the network-flow based *fixed-charge-matching* model. The equivalence of the MILP reformulation is also verified experimentally. For the existing MILP-based models, common notation is established and used to demonstrate that they are extensions of the base models in straight-forward ways. The contributions of this study are threefold. Firstly, it helps the analyst to understand the structural commonalities and differences of current conflation models and to choose different models. Secondly, by reformulating the network-flow models (and therefore, all current models) using MILP, the presented work eases the practical application of conflation by leveraging the many off-the-shelf MILP solvers. Thirdly, the base models can serve as a common ground for studying and writing new conflation models by allowing a modular and incremental way of model development.

Keywords: data fusion; conflation; optimization; geographic information systems

Citation: Lei, Z.; Yuan, Z.; Lei, T.L. On the Theoretical Link between Optimized Geospatial Conflation Models for Linear Features. *ISPRS Int. J. Geo-Inf.* **2024**, *13*, 310. https://doi.org/10.3390/ijgi13090310

Academic Editors: Eliseo Clementini and Wolfgang Kainz

Received: 16 June 2024
Revised: 24 August 2024
Accepted: 27 August 2024
Published: 29 August 2024

1. Introduction

In GIS(Geographic Information System) data production, planning, and many areas of spatial studies, it is often necessary to combine map data from different sources, including maps produced by different agencies and at different times. One of the main difficulties lies in identifying the corresponding geospatial features (objects) in different maps. This process, known as matching, is often error-prone and unreliable, and conventionally requires a large amount of human intervention and manual labor. Consequently, numerous research efforts have been devoted to developing reliable methods for computerized map matching and conflation over the past four decades [1,2]. Conflation has been applied in the production of different types of data such as administrative boundaries [3–6], point features such as gazetteers [7], and networks such as roads and rivers [8–12].

One of the earliest attempts at automated conflation is the conceptualization of the Map Assignment problem [13] in the 1980s. It is based on a classic crew scheduling model

called the "assignment problem" (see, e.g., [14]), which seeks a minimum cost plan to assign workers to jobs on a one-to-one basis. Simple as it is, it embodies a natural strategy for map matching: assigning geospatial features in one map to features in another map based on minimizing their total discrepancies (or "cost" in terms of the assignment problem). Although there are other approaches to conflation, the natural optimization perspective is inherited in numerous recent research endeavors due to its simple assumptions and expressive power. As will be discussed in the next section, there are three classes of optimized conflation models. The first is the assignment problem-based models. This was the main model in the literature until the late 2010s. Due to some of the limitations of the assignment problem, a new class of conflation models was developed circa 2020 based on the minimum cost network flow problem (network flow problem for short hereafter). These new models [11,12,15] were more flexible than the assignment problem due to the increased power of the underlying network flow model. More recently, a third type of conflation model called the topological conflation model appeared in the literature. Unlike conventional models, the topological conflation model can preserve certain topological relationships, such as connectivity or node-arc incidence relationships during the conflation process. To enforce these new requirements, neither the assignment problem nor the network flow problem has sufficient structural flexibility. Consequently, these more advanced models were formulated using more general mathematical programming languages involving Mixed Integer Linear Programming (MILP).

Given the many seemingly different conflation models, a natural question is how the models compare structurally and which model should be chosen given such knowledge. To address this issue, this study identifies the fundamental linkages between the major optimized conflation models in the literature. In particular, a base MILP model is presented, from which all other conflation models can be built by adding constraints and parameters.

The main goal of the paper is to help understand the multitude of optimized conflation models in the literature. By means of the common base, one can see the difference between models in terms of what is added to the base. The base model also allows all existing models to be expressed in the same mathematical language (MILP), and may facilitate the development of conflation models in a modular and incremental way.

In the remainder of this paper, Section 2 provides an in-depth analytical review of optimized conflation models in the literature, with a focus on their structures and properties. The Method section presents the base MILP model *base-matching* and shows that it is equivalent to the network flow-based *fixed-charge-matching* (*fc-matching*) model. Based on this equivalence, it is further demonstrated how existing optimized conflation models (network-flow or MILP-based) are related to or differ from the base model. It is demonstrated that most models can be formulated by extending the base model with additional constraints or objective function terms. In the Experiment section, the equivalence between the *base-matching* model and the network-flow based model is verified numerically. This article then concludes with a summary of the findings.

2. Geospatial Data Conflation Methods

Geospatial conflation involves various processes and stages. Roughly speaking, one can decompose the conflation problem into two main stages. In stage one, the similarity or spatial offset between geographic features is measured, which expresses the likelihood that an individual feature pair from two datasets should match. This is the similarity measurement stage. It is important in that a strong similarity measure can produce high scores for a corresponding features pair that represent the same object, and therefore, increase their likelihood of being matched. The second stage is match selection, in which, given all the similarity/distance metrics in stage one, one selects a set of feature pairs to conflate. Optionally, there is a third stage in which one merges the geometries or attribute sets of the matched feature pairs in stage two.

Among these two stages, the similarity measurement problem has been extensively studied in the literature. A classic similarity metric is the Hausdorff distance. Given two

GIS features A and B, it measures the maximum deviation of any point of $p \in A$ to feature B as a point set, and vice versa, the maximum deviation of any point in B to A as a set. Mathematically, the Hausdorff distance is defined as:

$$H(A, B) = \max\left(\min_{p \in A} d(p, B), \min_{q \in B} d(q, A)\right)$$

This definition is faithful and the greater the difference between A and B, the greater their distance metric. When the distance metric is zero (or close to zero), the two features must be the same. In addition to the Hausdorff distance, numerous other similarity metrics have been employed in the literature, involving spatial offset and separation [9], angular distance and orientation [10,12,16], shape [16,17], and topological measurements such as the degree of nodes. While these metrics differ in details, they embody the same principles as the Hausdorff distance: GIS features that are close in location or other aspects should have a low distance/dissimilarity value. The reader is referred to the reviews in [1,2,16] for detailed coverage of the (dis)similarity metrics. That being said, the remainder of this section is focused on the second stage, match selection.

2.1. Heuristic vs. Optimized Conflation Methods

The simplest match selection method is based on a greedy strategy of matching the closest or most similar features. Depending on the order of data processing, there are a number of variant strategies. For example, one can sequentially match each feature in one dataset to its closest candidate feature in the other dataset. This strategy is the nearest neighbor join method and can be directly carried out using most existing GIS packages. Such spatial join strategies, as pointed out in [18], could lead to logically inconsistent matches. The closest relationship may not be reflexive. That is, the closest feature in J for a feature $i \in I$ may have the closest feature in I that is different from I.

There are different ways to address conflicts in match results. For example, one could select the pair of closest features to match and then exclude them in future steps. This is called the k-Closest Pair Query (KCPQ) [19] and is widely used in database research. Beeri et al. [18] proposed a two-sided nearest neighbor join method, in which they choose among possibly conflicting pairs of candidate matches of a local area using a probabilistic score between 0 and 1. This is followed by subsequent studies by Tong et al. [20]. Another method for coping with map conflicts is to eliminate them in advance. This is exemplified by a general technique called the Rubbersheeting process, which dates back to the early days of conflation research [3,4]. The general idea is to reduce the spatial displacement between two maps by iteratively identifying and merging control points (called "anchors") on the two maps, and then applying a continuous transformation (e.g., affine) to the local regions between these anchors. The anchor points are typically points that are easily identifiable by the human expert (road junctions, land marks, etc.). After the spatial displacement is reduced, a simpler method, such as spatial join and polygon overlay, is used for match selection.

Generally, the aforementioned methods are greedy and sequential in nature [9,10]. They make match selections one by one and cannot undo an erroneous match once it is made. In comparison, a different type of match selection method, called optimization-based conflation method, does not suffer from these issues. As will be discussed shortly, they match all features "simultaneously" by treating the match selection as an optimization problem of minimizing the total discrepancy between two maps.

The heuristic methods are fundamentally different from the optimization-based methods in that the match solution for such an algorithm is not unique in nature. For example, the match produced by the nearest neighbor join (i.e., greedy) method may well depend on the order in which the data are processed. It even depends on which dataset is used as the source dataset and which is used as the target (see [18]). In contrast, any optimization based conflation model has a unique solution with respect to the pre-specified match condition: the optimal solution. Therefore, one cannot generally express a heuristic conflation method as an optimized conflation model.

2.2. Existing Optimized Conflation Models

This subsection presents an analytical review of the main optimized conflation models in the literature, starting with the map assignment problem in the 1980s. This includes the formulation of the optimization problem for each conflation model, as well as a discussion of its functional features and structural characteristics. The models presented include (1) the original assignment problem and a set of common notation for the assignment problem and all subsequent models, (2) the network-flow based fixed-charged matching problems which offer more structural flexibility, (3) a unified bidirectional matching model aimed at reconciling m:1 (and 1:n) matches in the two opposite directions of match, (4) a topological conflation model aimed at preserving edge-to-edge connectivity during the match, and (5) a node-arc topological model aimed at preserving the node-arc incidence relation during the matching.

2.2.1. Common Notation

Throughout this paper, MILP is used as a common language to express new and existing optimized conflation models. Generally speaking, MILP is a high-level algebraic modeling language in which the decision problem (i.e., choice of match relation in this context) is expressed in terms of sets and decision variables. The sets contain constants/parameters describing the objects involved in the optimization model and the attributes of these objects, while the decision variables describe the actions/outcome of the model (e.g., which objects are matched). The MILP model itself is in essence a set of linear equations about the requirements that the final solution (e.g., match relation) must satisfy.

For consistency, a set of common definitions for both the constants and variables are presented below, which will apply to all conflation models in this article. Variables and parameters in the original articles of conflation models may be renamed, if necessary, to conform to the common notation defined here.

I, J are the two geospatial datasets to be matched and conflated. d_{ij} is a directed distance or dissimilarity metric (such as the directed Hausdorff distance) that supports membership relation decisions. The directed distance d_{ij} is zero if the feature $i \in I$ coincides with a part of the feature $j \in J$. $d\prime_{ij}$ is the same distance/dissimilarity metric measured in the opposite direction ($J \to I$) $D_{ij} = max(d_{ij}, d\prime_{ij})$ is the total distance, defined as the greater of the two directional distances. c is a cutoff distance, beyond which two features are considered too distant/dissimilar to be a potential match. M is a sufficiently large number that ensures that all coefficients of the form $M - \cdots$ in a given model are positive.

With the above distance definitions, one can define the sets of potential matches:

$F = \{(i,j) | d_{ij} < c,\ i \in I, j \in J\}$ is the set of potential forward partial matches from I to J where the directed distance d_{ij} is less than the cutoff distance c.

$G = \{(i,j) | d\prime_{ij} < c,\ i \in I, j \in J\}$ is the set of potential partial backward matches from features in J to features in I

$E = \{(i,j) | D_{ij} < c,\ i \in I, j \in J\}$ is the set of potential partial full matches between features of I and J.

The decision variables are:

$x_{ij} = 1$ if feature i is considered the same object as j, and 0 otherwise.

$y_{ij} = 1$ if feature $i \in I$ corresponds to part of feature $j \in J$, and 0 otherwise.

$z_{ij} = 1$ if feature $j \in J$ corresponds to part of $i \in I$, and 0 otherwise.

2.2.2. The Assignment Problem (AP)

The Map Assignment problem [13], is the first optimized conflation model in the literature. As mentioned in Section 1, it is inspired by the classic job assignment problem in operations research for assigning a set of n workers to the same number of jobs. The goal is to find a minimum cost match plan (in terms of time) while respecting the 1:1 relation between the two sets. In the map assignment problem, the distance/dissimilarity

between two features has been used as the assignment cost (instead of time). Using the aforementioned common notation, the assignment problem can be expressed as:

$$\text{minimize} \sum_{i \in I, j \in J} d_{ij} \cdot x_{ij} \tag{1}$$

Subject to:

$$\sum_{j \in J} x_{ij} = 1 \; \forall i \in I \tag{2}$$

$$\sum_{i \in I} x_{ij} = 1 \; \forall j \in J \tag{3}$$

The objective Function (1) is aimed at minimizing the total matching distance. It is expressed as the sum of distances $\sum_{i \in I, j \in J} d_{ij} \cdot x_{ij}$ for the matches that have actually been selected (those for which $x_{ij} = 1$). Constraints (2) and (3) maintain that the match relation x_{ij} is one-to-one. That is, each feature $i \in I$ is matched to exactly one feature in J (2) and vice versa, each feature in J is matched to exactly one in I (3). As the constraints specify the number of features that can be/must be assigned to each feature in I ($\sum_{j \in J} x_{ij}$) and each feature in J ($\sum_{j \in J} x_{ij}$), they are called the "cardinality" constraints. In the assignment problem, the cardinality constraints are the only constraints.

Li and Goodchild [8] were the first to implement and test the map assignment problem. They found that one of the two cardinality constraints may not be feasible because the two datasets I and J are typically not equal in size. Assuming that the size of I is smaller than that of J, they changed constraints (3) to the following inequality form:

$$\sum_{i \in I} x_{ij} \leq 1 \; \forall j \in J \tag{4}$$

Li and Goodchild [9] extended the basic assignment model by considering partial matches. To this end, they used the directed Hausdorff distance instead of the full Hausdorff distance to measure distance. Being a half distance, the directed Hausdorff distance is zero when one feature coincides with part of a target feature. Additionally, they defined and used a cutoff distance c, and defined the similarity of a feature $i \in I$ to a feature $j \in J$ as follows:

$$s_{ij} = \begin{cases} 0 & \text{if } d_{ij} > c \\ c - d_{ij} & \text{otherwise} \end{cases}$$

where d_{ij} denotes the directed Hausdorff distance. They then defined a model as follows (with a slight change in notation):

$$\text{maximize} \sum_{i \in I} \sum_{j \in J} s_{ij} \cdot y_{ij} \tag{5}$$

s.t.

$$\sum_{j \in J} y_{ij} \leq 1, \; \forall i \in I \tag{6}$$

$$\sum_{i \in I} l_i y_{ij} \leq \alpha l_j, \text{ for each } j \in J \tag{7}$$

$$\sum_{i \in I} y_{ij} + \delta_j \geq 1, \text{ for each } j \in J \tag{8}$$

where l_i, l_j are the lengths of features (roads in [9]). δ_j is a parameter defined to be 1 if j has no nearby feature (within the cutoff distance c), and zero otherwise.

Constraints (6) are similar to the cardinality constraints (4) of the assignment problem. Constraints (7) maintain that the total length of the lines assigned to a target line should not exceed the target's length. α is a parameter used to allow for errors in the lengths of the involved features. Constraints (8) maintain that if there are nearby features to j (i.e., $\delta_j = 1$),

then one of these nearby features *must be assigned* to j. Li et al. [9] defined an analogous model for the reverse direction of matching from J to I. They called these two models sub-model 1 and sub-model 2. They then applied the two sub-models one by one and solved any inconsistencies between the two sub-models in a post-processing procedure. It should be noted that the constraint sets (7) and (8) could lead to infeasibility. In some occasions, constraints (8) could be used to force one feature $i \in I$ to be assigned to two different target features j and j' (therefore violating (6)).

The assignment problem formulation was followed in the subsequent research. For example, Tong et al. [10] applied the original assignment model to their road network data and reported a low match rate of 56.5%. Although simple in structure, the assignment problem formulation has its limitations. It has a strong requirement that all features in (one of) the datasets must be assigned (constraint (2)). Li et al. [9]'s work, while enhancing the assignment problem in several directions, introduced its own issues. Such issues are further discussed in [15].

2.2.3. The Fixed-Charge Matching (fc-Matching) Problems

The fixed-charge matching (fc-matching) problems represent an improvement over the assignment problem with less stringent requirements and a more flexible structure. This is achieved by replacing the underlying optimization model in the assignment problem with a more powerful model called the minimum cost network-flow circulation problem (network flow problems for short).

To address the limitations of the assignment problem formulation, Lei et al. [11,15] proposed two new optimized conflation models based on the network flow problem. The network flow problem is a more powerful and flexible model than the assignment problem as it can express a range of optimization problems including the shortest path, the assignment problem itself, and the fixed-charge matching problems, among many others. One advantage of the network-flow based conflation models is that they can be solved using fast specialized algorithms such as the push-relabel algorithm. However, the problem format of the network flow is different from the commonly used MILP models and may require a separate optimization package to solve.

The fixed-charge matching problems, first defined in [11] include two models called the *fc-matching* and the *fc-bimatching* models. They differ in that the *fc-matching* model assumes a one-to-one correspondence between the two matched GIS datasets, whereas the *fc-bimatching* model allows many-to-one (and one-to-many) correspondence. The fixed-charge-matching models are special instances of the network problem. The network problem by definition, optimizes the number of flows along the edges of a specially designed network. Each edge in the network is directed and has an associated cost for carrying a unit amount of flow, an upper bound, and a lower bound for the flow. The only requirement is *flow preservation* at each node. That is, the amount of flow entering a network node must be equal to the amount of flow leaving that node. The decision variables are the number of flows f_e along the network edges e, and the system objective is to minimize the total flow cost. In the context of matching, the amount of flow represents the possible matching of an object i in one dataset I to an object j in a second dataset J. It is 1 if i is assigned to j and 0 otherwise.

Figure 1 presents a diagram for the one-to-one *fc-matching* problem. The labels on the network edges represent their attributes: the flow cost, the lower bound, and the upper bound. Most edges have a lower bound of 0 and an upper bound of 1 (representing an assignment or non-assignment). On these occasions, the lower and upper bounds are omitted and only the flow cost is shown on the label. In this network, all flows come out of the source node s and eventually enter the sink node t, except for the rebalancing flows from t to s. Each node in dataset I is linked to the source, and each node in J is linked to the sink with zero-cost edges. The upper bound of 1 on these edges ensures that each $i \in I$ (and $j \in J$) can be assigned to at most one node in the other dataset. The links between the nodes of I and nodes of J represent the actual assignment, and the edge costs represent the

distance/dissimilarity metrics. All these costs are positive, making the model seeking a minimum cost match plan. However, the cost of the rebalancing link *ts* is set to a negative value ($-F$) to ensure that some non-zero flow will be generated in the network.

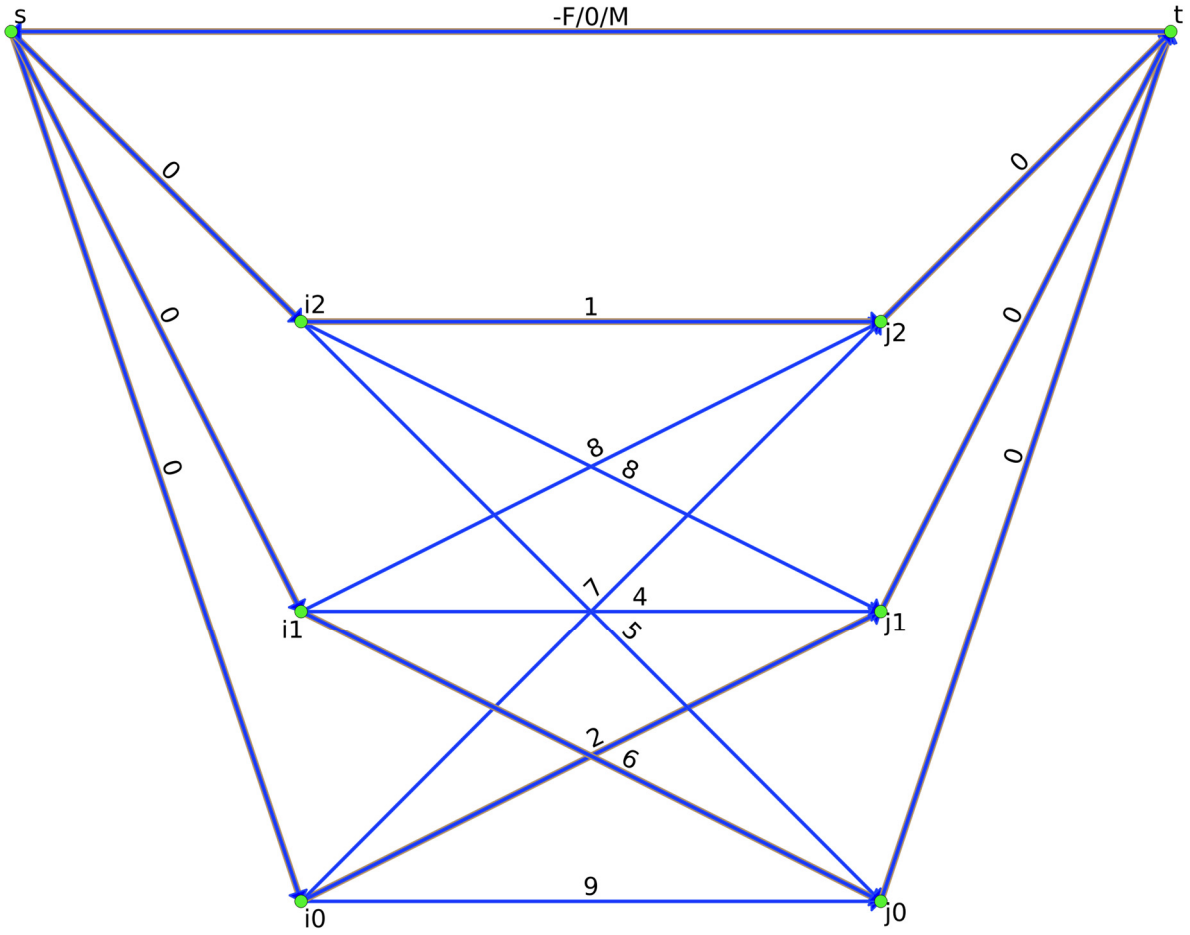

Figure 1. The flow network for the fc−matching problem. Adapted from [11].

Figure 2 presents a schematic view of the *fc-bimatching* model, structured as a min-cost network flow problem. Structurally, the network roughly consists of two halves. The left sub-network represents the assignment of objects from dataset I ($i_0, i_1, \cdots$) to objects in the dataset J. The link from a node in i to a node in j represents a *possible* assignment from object i in dataset I to object j in J (e.g., from i_1 to j_2). The right sub-network represents possible assignments in the opposite direction from J to I. As in the *fc-matching* diagram, the labels on the network edges represent their flow cost and lower/upper bounds.

There are several special source and sink nodes: s, t, r, o, a, b, e. They inject flows to and absorb flows from the regular nodes in I and J. Links between these special nodes have a lower bound of 0 and an upper bound of infinity (or a very large number M). Most of them have zero flow costs. The only exceptions are the flow rebalancing link $b \rightarrow a$ and the excess flow link $e \rightarrow b$. The rebalancing link has a negative cost ($-F$) and provides an incentive for the flow network to generate non-zero flows (assignments). In this study, F is set to be the cutoff distance c plus one. The excess flow link has a positive cost P for penalizing excessive multi-assignments to one object. This is because multi-assignments represent partial matches, which are harder to characterize in terms of their match relations. For example, in the one-to-one *fc-matching* problem, any feature can be matched to only one target feature. This means that the match is exclusive. Once matched, a feature cannot be involved in any other match. This property helps in reducing ambiguous and erroneous matches. In comparison, no such constraints can be imposed in the many-to-one

fc-bimatching problem. By the definition of many-to-one matching, many features may be matched to one target. If a line is split into three parts, nothing stops the model from matching one or two of the parts to the correct target feature while mismatching the other parts. Therefore, partial matches are more error-prone. This is why the penalty factor P is imposed to discourage multi-assignments.

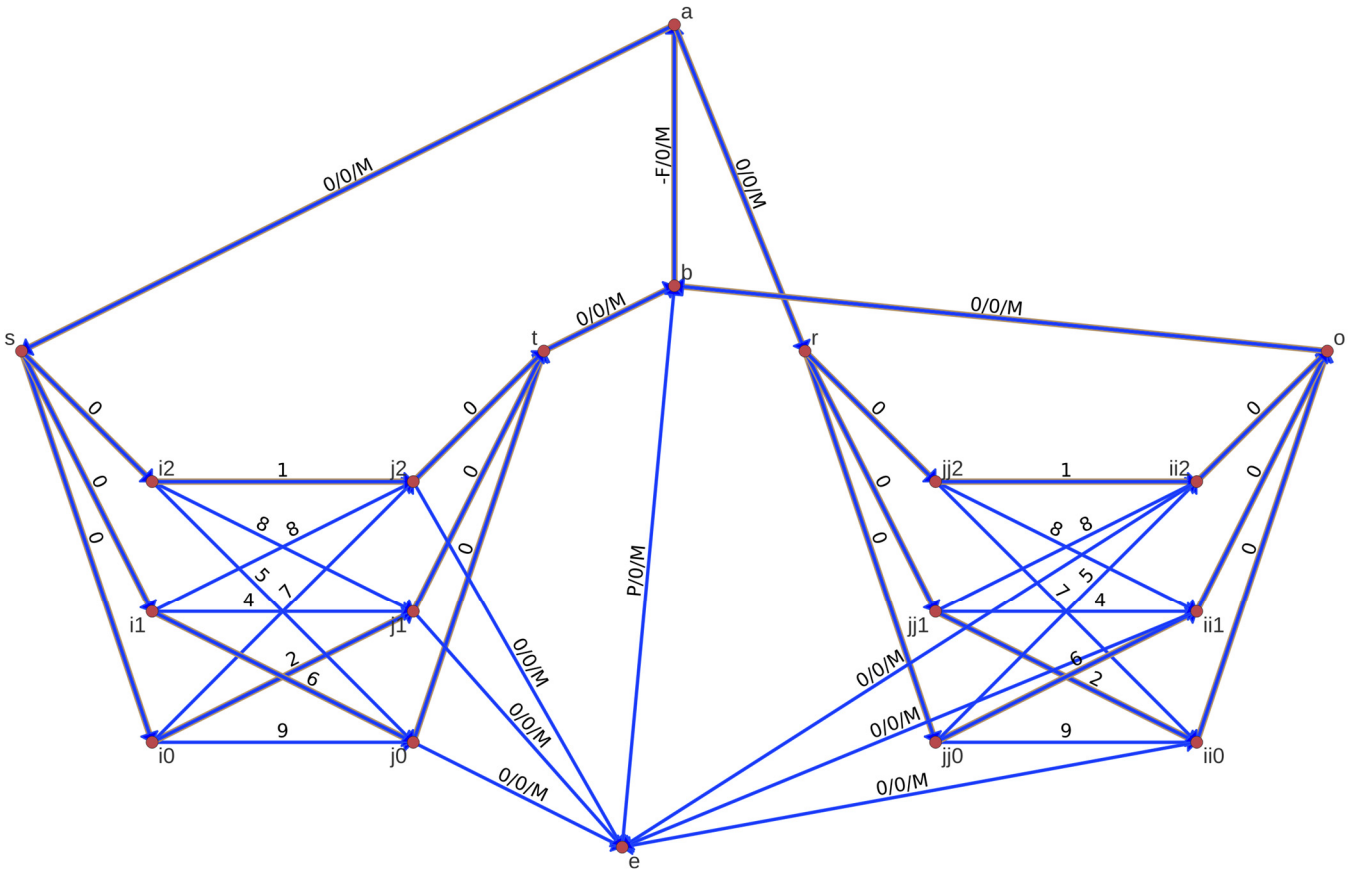

Figure 2. The flow network for the fc−bimatching problem. Adapted from [11].

Two earlier versions of the fixed-charge matching problems called the *p-matching* problems, were defined in [15]. They are similar in structure to the *fc-matching* models, except that they pre-define the number of matches to p. This is enforced by setting both the upper and lower bounds of the flow rebalancing link to p. The *p-matching* requires a search for an appropriate value of p based on the level of matching errors at each p. The *fc-matching* models improve the *p-matching* model by allowing p to be automatically determined by optimization.

Wu et al. [12] developed a modified version of the network flow based road conflation model, which is similar in structure to the *fc-matching* problem. However, they defined the capacity of an edge ij based on the lengths of the two roads being matched. More specifically, they defined the capacity for edge $(i, j) \in E$ to be: $l_{ij} = max(len_i, len_j)$, where len_i, len_j are the lengths of i and j. Additionally, they used a hypothetical null object in each dataset to represent non-matches. They also used a pre-processing step to preclude unlikely edges from the flow networks based on a number of characteristics of the pair of candidate roads. This includes their angle, length capacity (similar to (7)), etc. Their experimental results showed that these modifications reportedly improved the accuracy of matching compared with *fc-matching*.

2.2.4. The Unified Bidirectional Matching u-Bimatching Problem

The main purpose of the unified bidirectional matching problem [21] is to eliminate potentially conflicting matches between the two opposite match directions. As discussed earlier, these inconsistencies between the two match directions may occur in m:1 matching models.

In the *fc-bimatching* model, the partial assignments (i.e., m:1 matches) are penalized in favor of full, one-to-one assignments. When possible, the model attempts to make full assignments first. However, the *fc-bimatching* model does not preclude inconsistent partial assignments. To address this issue, Lei et al. [21] attempted to eliminate erroneous partial matches by imposing new constraints in optimized conflation models. They achieved this by including full and partial matches in a new optimization model with a set of "link" constraints to preclude inconsistent assignments. Using the common notation defined at the beginning of this section, their model called the unified bidirectional matching model *u-bimatching*, can be described in MILP as follows:

$$maximize\ Z = \alpha \cdot \sum_{(i,j)\in F\cap G} B_{ij}x_{ij} + \sum_{(i,j)\in F} b_{ij}y_{ij} + \sum_{(i,j)\in G} b\prime_{ij}z_{ij} \tag{9}$$

Subject to:

$$\sum_{(i,j)\in F\cap G} x_{ij} \le 1,\ \text{for each } i \in I \tag{10}$$

$$\sum_{(i,j)\in F\cap G} x_{ij} \le 1,\ \text{for each } j \in J \tag{11}$$

$$\sum_{(i,j)\in F} y_{ij} \le 1,\ \text{for each } i \in I \tag{12}$$

$$\sum_{(i,j)\in G} z_{ij} \le 1,\ \text{for each } j \in J \tag{13}$$

$$x_{ij} + y_{ij} \le 1,\ \text{for each } (i,j) \in F \cap G \tag{14}$$

$$x_{ij} + z_{ij} \le 1,\ \text{for each } (i,j) \in F \cap G \tag{15}$$

$$N \cdot y_{ij} + \sum_{(k,j)\in F\cap G} x_{kj} + \sum_{(i,l)\in F\cap G} x_{il} + \sum_{(k,j)\in G, k\neq i} z_{kj} \le N,\ \text{for each}(i,j) \in F \cap G \tag{16}$$

$$N \cdot z_{ij} + \sum_{(k,j)\in F\cap G} x_{kj} + \sum_{(i,l)\in F\cap G} x_{il} + \sum_{(i,l)\in G, l\neq j} y_{il} \le N,\ \text{for each}(i,j) \in F \cap G \tag{17}$$

In the above, N is a sufficiently large number for formulating certain constraints such as (16) and (17) (in the so-called "big M" method in operations research [14]). α is a relative weight value for emphasizing full assignments x_{ij} (as compared to partial assignments, $\alpha \ge 2$).

In the *u-bimatching* formulation, the objective function maximizes the benefits B_{ij}, b_{ij}, $b\prime_{ij}$ associated with making full, forward, and backward partial assignments, respectively. Constraints (10) and (11) are cardinality constraints stating that each feature in I can be assigned to at most one target feature in J, and vice versa each $j \in J$ can be assigned to at most one $i \in I$. They establish a one-to-one correspondence with the decision variable x_{ij}. Constraints (12) and (13) are the cardinality constraints for the forward partial assignments y_{ij} and backward partial assignments z_{ij}, respectively. Either establish unilaterally that a source feature can "belong" to at most one target feature. Constraints (14) and (15) maintain that a source feature can not be assigned to a target both as a full assignment and as a partial assignment.

Constraint (16) is a link constraint that ensures the compatibility of the forward/backward partial assignments as well as the full assignments. The left-hand-side (LHS) has four terms. Clearly, if any of the last three terms are positive, then it forces y_{ij} to be zero. That is, i cannot be assigned as a part of j. If all of them are zero, then y_{ij} is allowed

to be one. The three terms represent three occasions in which the assignment y_{ij} should not happen. This includes: (a) if j is assigned in a full match, (b) if i is assigned in a full match, and (c) j is assigned to some feature $k \in I$. Obviously, the first two cases are incompatible with $y_{ij} = 1$. The last case (c) is also incompatible because it forms a chain of transitive assignments $i \to j \to k$, implying that $i \in I$ is part of a different feature $k \in I$, which is absurd. Constraint (17) is similar to constraint (16), except that the roles of I and J are swapped.

2.2.5. The Edge Connectivity Based Matching (ec-Matching) Problem

The edge connectivity based matching (ec-matching) problems (introduced in [22]) are an extension of the previous conflation models in that they can ensure that the edge-to-edge connectivity relation is preserved between the two matched datasets (as it should be). The conflation models reviewed so far match the features of the two datasets at the individual feature level. Certain consistency conditions have been enforced. This includes, for example: cardinality constraints (one feature may not be assigned to two different targets), consistency between opposite assignments (transitive assignment is not allowed), and the penalty on multi-assignments. However, no attention has been paid to the consistency between inter-feature relationships, such as adjacency and incident relations. Such relationships are often used by human experts to match difficult cases. For example, when there are large spatial offsets, one may need to check nearby neighbors to determine whether a pair of features is a true match. If the neighbors all match, then it is likely that they should match as well. Such "tracing" reflects the fact that human experts use topological relations during conflation. By analogy, conflation models should also respect the topological relations so that they are the same on either side of the match. For linear networks, there are at least two types of topological relations: edge connectivity and edge-node incident relations. Two types of optimized conflation models have been developed in the literature.

The first type of topological relation is edge connectivity. If two roads are connected in reality, they should be so in both GIS datasets. Consequently, it would be an error to match a pair of connected roads in I to a pair of disconnected roads in J. Based on this notion of preserving connectivity, Lei et al. [22] proposed the edge connectivity based matching *ec-matching* problem. In addition to the common notation established earlier, the following are needed:

$r_{ik} = 1$ if $i, k \in I, k \neq i$ are connected, and 0 otherwise.
$t_{jl} = 1$ if $j, l \in J, j \neq l$ are connected, and 0 otherwise.

$$P_{ijk} = \left\{ l \in J \middle| (k,l) \in E, t_{jl} = 0 \right\}, \text{ for each } (i,j) \in E, k \in I, r_{ik} = 1.$$
$$Q_{ijl} = \{ k \in I | (k,l) \in E, r_{ik} = 0 \}, \text{ for each } (i,j) \in E, l \in J, t_{jl} = 1$$

P_{ijk} is a (forward) incompatibility set containing all the edges $l \in J$ for which i, k are connected in I and j, l are *not* connected in J. For the aforementioned reasons for topological consistency, one can see that if the assignment x_{ij} is made, then the assignment x_{kl} cannot happen. Similarly, Q_{ijl} is a backward incompatibility set containing the set of elements in I that cannot be assigned to l if the assignment $x_{ij} = 1$.

With the addition notation, the *ec-matching* model [22] is formulated in MILP as:

$$maximize\ Z = \sum_{(i,j) \in E} B_{ij} x_{ij} \tag{18}$$

Subject to:

$$\sum_{(i,j) \in E} x_{ij} \leq 1, \text{ for each } i \in I \tag{19}$$

$$\sum_{(i,j) \in E} x_{ij} \leq 1, \text{ for each } j \in J \tag{20}$$

$$\left|P_{ijk}\right| \cdot x_{ij} + \sum_{l \in P_{ijk}} x_{kl} \leq \left|P_{ijk}\right|, \text{ for each } (i,j) \in E, k \in I, r_{ik} = 1 \tag{21}$$

$$\left|Q_{ijl}\right| \cdot x_{ij} + \sum_{k \in Q_{ijl}} x_{kl} \leq \left|Q_{ijl}\right|, \text{ for each } (i,j) \in E, l \in J, t_{jl} = 1 \tag{22}$$

Constraints (19) and (20) are cardinality constraints. Constraint (21) embodies the aforementioned connectivity preservation idea. For a pair of candidate features i, j, if any assignment in the incompatibility set P_{ijk} happened, then x_{ij} is forced to be zero. By enumerating all possible incompatible sets, the model makes it impossible to match a connected pair $i, k \in I$ to a disconnected pair $j, l \in J$. Symmetrically, constraint (22) forbids any connected pair in J to be matched to disconnected pairs in I.

2.2.6. The m:1 (1:n) Element Connectivity Bi-Matching Problem (ec-Bimatching)

Lei et al. [22] also proposed a many-to-one version of the connectivity based conflation model, called the *ec-bimatching* problem. The model is similar in structure to the *ec-matching* model except that it deals with m:1 (and 1:n) matches. Therefore, directed distances are used instead of full distances. Accordingly, the two incompatibility sets are modified to the following forms:

$$A_{ijk} = \left\{ l \in J \middle| (k,l) \in F, t_{jl} = 0 \right\}, \text{ for each } (i,j) \in F, k \in I, r_{ik} = 1$$

$$B_{ijl} = \{ k \in I | (k,l) \in G, r_{ik} = 0 \}, \text{ for each } (i,j) \in G, l \in J, t_{jl} = 1$$

With the modified incompatibility sets and partial assignment variables, the *ec-bimatching* problem [22] can be formulated in MILP as:

$$maximize\ Z = \sum_{(i,j) \in F} b_{ij} y_{ij} + \sum_{(i,j) \in G} b\prime_{ij} z_{ij} \tag{23}$$

Subject to:

$$\sum_{(i,j) \in F} y_{ij} \leq 1, \text{ for each } i \in I \tag{24}$$

$$\sum_{(i,j) \in G} z_{ij} \leq 1, \text{ for each } j \in J \tag{25}$$

$$\left|A_{ijk}\right| \cdot y_{ij} + \sum_{l \in A_{ijk}} y_{kl} \leq \left|A_{ijk}\right|, \text{ for each} (i,j) \in F, k \in I, r_{ik} = 1 \tag{26}$$

$$\left|B_{ijl}\right| \cdot z_{ij} + \sum_{k \in B_{ijl}} z_{kl} \leq \left|B_{ijl}\right|, \text{ for each} (i,j) \in G, l \in J, t_{jl} = 1 \tag{27}$$

The constraints above are similar to those of the *ec-matching* model, except for the use of partial assignment variables. The original formulation of [22] included link constraints of a form similar to (16) and (17), which are omitted here for simplicity and ease of comparison.

2.2.7. The Edge-Node Matching (en-Matching) Problem

The edge-node matching (en-matching) problem [23] introduces a second type of topological condition, namely, the node arc incidence relationship, into optimized conflation. As an example of such conditions, if a road $i \in I$ is matched to a road $j \in J$, then the end nodes of i and j as road junctions must match. Otherwise, the match relation is topologically incorrect, and should not be made. To preserve the node-arc relation during matching, the nodes and the edges of the two linear networks must be matched simultaneously and consistently. To this end, Lei et al. [23] proposed an edge-node matching (*en-matching*) model, which requires additional definitions as follows:

Let $V(I)$ be the vertex set of I, and $V(J)$ be vertex set of J,

$N = \{(r,s)|d(r,s) < c,\ r \in V(I), s \in V(J)\}$ is the set of candidate node pairs whose distances are less than the cutoff distance.

Given an edge in I (or J), let $f(i)$ denote its from-node, and $t(i)$ denote its to-node.

$\overline{D}_{rs}$ is a distance metric between the nodes in $V(I)$ and $V(J)$.

$\overline{b}_{rs} = M - \overline{D}_{rs}$ is the benefit or similarity metric for matching nodes r and s.

β is a weight value for making nodal matches (vs. making edge matches). Lei et al. [23] assumed $\beta = 4$ based on the assumption that each junction is associated with four roads on average. Lei et al. [23] also used a parameter γ to prioritize the matching of higher-degree nodes. This is to avoid isolated "islands" of matched cliques (see [23] for details).

A new decision variable for matching two nodes is needed:

$u_{rs} = 1$ if $r \in V(I)$ is assigned to $s \in V(J)$, and 0 otherwise.

Now the *en-matching* problem can be formulated in MILP as:

$$Maximize Z = \sum_{(i,j)\in E} \left(M - D_{ij} + \gamma \cdot M\right) x_{ij} + \beta \cdot \sum_{(r,s)\in N} \left(M - D_{rs}\right) u_{rs} \tag{28}$$

Subject to:

$$\sum_{(i,j)\in E} x_{ij} \leq 1 \text{ for each } i \in I \tag{29}$$

$$\sum_{(i,j)\in E} x_{ij} \leq 1 \text{ for each } j \in J \tag{30}$$

$$\sum_{s\in M} u_{rs} \leq 1 \text{ for each } r \in V(I) \tag{31}$$

$$\sum_{r\in N} u_{rs} \leq 1 \text{ for each } s \in V(J) \tag{32}$$

$$u_{f(i)f(j)} + u_{f(i)t(j)} \geq x_{ij} \text{ for each } (i,j) \in E \tag{33}$$

$$u_{f(i)f(j)} + u_{t(i)f(j)} \geq x_{ij} \text{ for each } (i,j) \in E \tag{34}$$

$$u_{t(i)t(j)} + u_{f(i)t(j)} \geq x_{ij} \text{ for each } (i,j) \in E \tag{35}$$

$$u_{t(i)t(j)} + u_{t(i)f(j)} \geq x_{ij} \text{ for each } (i,j) \in E \tag{36}$$

$$x_{ij} \in \{0,1\} \text{ for each } (i,j) \in E \tag{37}$$

$$u_{rs} \in \{0,1\} \text{ for each } (r,s) \in N \tag{38}$$

In the above, the objective function (28) maximizes the total weighted similarity between matched edges and between matched nodes. Constraints (29) through (32) are the cardinality constraints for the nodal and edge assignments, as before. Constraint (33) is one of the topological constraints. It states that if the edge assignment x_{ij} is made, then its from-node $f(i)$ should be matched to either the from-node of j or the to-node of j. Constraints (34), (35), and (36) express similar conditions. Collectively, constraints (33) through (36) maintain that if i is matched to j, then their from- and to-nodes must match.

In the existing literature, most optimized conflation methods are developed for linear features such as roads [9–11,20,24–27]. This is probably due to the importance of roads as a spatial reference and their use in early studies by the US Census. By comparison, optimized conflation for point and polygon features received less research attention. For point features, Li and Goodchild's work [8] used a hypothetical point dataset (along with a road dataset) to compare the optimization-based assignment problem and two greedy matching algorithms for conflation. However, it was not followed by others in the literature. This is probably due to the fact that point features are relatively simple in structure. Conflation for polygon features is pre-dominated by heuristic methods. A common example is the polygon overlay method (see e.g., [28,29]) in which two polygon features from different sources are considered the same if their geometric intersection is

large enough (compared to the two original features). To the best of our knowledge, no optimized conflation studies have been published for polygon data so far.

2.2.8. M:N Matching Methods

So far, this article has considered the optimized conflation model in the more strict sense of [9,11]. That is, only models that can find the optimal match plan (i.e., minimum distance match) are discussed. In the literature, some scholars (see e.g., [26]) also consider a broader class of conflation algorithms as optimization based, in a weaker sense that they reduce the distance or some other metrics (without necessarily finding the optimal solution). In this broader category, some of the conflation algorithms can consider the more complex m:n match case.

For example, the "heuristic probabilistic relaxation" method [24,30] can take into account spatial context and therefore, can handle the m:n match cases. In [30], a confidence matrix is computed between pairs of road intersections using relative distances. Then, repeatedly, joint compatibility of two neighboring features is computed and used to update the confidence of matches until the confidence level is sufficiently high. Yang et al. [24] extended the work of [30] to match road features.

Fu et al. [31] adopted a multi-variable logistic regression approach for conflation, in which the Hausdorff distance, string distance, and the direction difference of two roads were used to classify candidate road pairs into two classes: match and unmatch. Guo et al. [26] extended logistic regression based conflation by using the so-called "strokes" (or sequences of road segments) as the candidates for matching. Consequently, their method can also handle the m:n matching case.

It should be noted that the heuristic probabilistic relaxation and logistic regression methods are heuristic in nature, and cannot guarantee to find the minimum discrepancy match plan as the mathematical programming based models reviewed in the previous subsections. Nonetheless, they are briefly discussed here due to their similarity to (strictly) optimized conflation models.

3. Method

From the previous section, one can see that there are many different optimized conflation models, starting from the simple assignment problem to the complex topological conflation models. In this section, a base model, formulated in MILP, is presented as a common foundation for all the above models. The base model, including a 1:1 version and a m:1 version, can be used to build the other optimized conflation models in the literature by appropriate transformations or the addition of constraints. It is also demonstrated in this section that the 1:1 version of the base model is equivalent to the *fc-matching* problem. At the end of the section, Table 1 summarizes the connection between each existing model and the base models.

Table 1. Link between existing conflation models and the base models.

Model	Base Model	Added Constraints	Additional Modifications
Assignment	base-matching	set $c = \infty$	input I, J are filtered by c as in Obs. 1
fc-matching	base-matching	None	
fc-bimatching	base-bimatching	Equations (52)–(55)	added penalty term in objective
u-bimatching	base-(bi)matching	Equations (14)–(17)	two base models are merged with priority to full assignments
ec-matching	base-matching	Equations (21)–(22)	
ec-bimatching	base-bimatching	Equations (26)–(27)	
en-matching	base-matching	Equations (33)–(36)	base model is replicated for nodes and arcs

3.1. A Base MILP Model

Let $B_{ij} = M - D_{ij}$ be the similarity measure (or "benefit") associated with making the full match x_{ij}.

The common base model (called *base-matching*) can be formulated in MILP using the common notation established in Section 2.2.1 as follows:

$$maximize\ Z = \sum_{(i,j)\in E} B_{ij}x_{ij} \tag{39}$$

Subject to:

$$\sum_{(i,j)\in E} x_{ij} \leq 1,\ \text{for each } i \in I \tag{40}$$

$$\sum_{(i,j)\in E} x_{ij} \leq 1,\ \text{for each } j \in J \tag{41}$$

The base model maximizes the total benefit in the objective function (39) and the only constraints are the two cardinality constraints (40) and (41) stating that the total number of assignments from each feature in I (or to each feature in J) can be at most one. Together, they enforce the 1:1 matching relation between I and J.

3.1.1. Equivalence to fc-Matching

It can be shown that the base model above is equivalent to the *fc-matching* problem described in Section 2 if one sets M in the base model as the fixed charge F in *fc-matching*. The proof is as follows:

Observation 1. *The base model (39) through (41) generates the same optimal solutions as the fc-matching problem.*

Proof. Note that there is a one-to-one correspondence between the x_{ij} variable in the base model and the network edge leading from i to j in the network flow diagram of *fc-matching* in Figure 1.

Let $I' = \{i|(i,j) \in E\}$ and $J' = \{j|(i,j) \in E\}$, and suppose that without loss of generality, $|I'| \leq |J'|$. At optimality, one must have

$$\sum_{(i,j)\in E} x_{ij} = 1,\ \text{for each } i \in I'$$

There should be exactly $|I'|$ assignments in the optimal base solution. This is because one can always make additional assignments to get more benefit in the objective if the number of matches is less than $|I'|$. On the other hand, there cannot be more than $|I'|$ assignments as it will break the cardinality constraint.

One can rewrite the objective as

$$maximize\ Z = M \cdot \sum_{(i,j)\in E} x_{ij} - \sum_{(i,j)\in E} D_{ij}x_{ij}$$

Since at optimality the first term:

$$\sum_{(i,j)\in E} x_{ij} = |I\prime|$$

is a constant, the base model is equivalent to:

$$minimize\ Z = \sum_{(i,j)\in E} D_{ij}x_{ij} \tag{42}$$

Subject to:

$$\sum_{j \in J'} x_{ij} = 1, \text{ for each } i \in I\prime \tag{43}$$

$$\sum_{i \in I\prime} x_{ij} \leq 1, \text{ for each } j \in J\prime \tag{44}$$

This model can be viewed as a restricted assignment problem instance defined on I' and J'.

Likewise, for the *fc-matching* problem in Figure 1, at optimality, the left column of nodes corresponds to I', and every node in I' must have a non-zero outgoing flow (again assuming $|I'| \leq |J'|$). There will be exactly $|I'|$ non-zero outgoing flows due to the defined edge capacities. In addition, at the flow rebalancing link, there will be exactly $|I'|$ amount of flow, each costing $-M$, a constant. Therefore, the optimal solution value is determined by the minimum cost assignments between I' and J'. This is also equivalent to the restricted assignment problem in (42) through (44).

Since both the *base-matching* and the *fc-matching* models are equivalent to the restricted assignment problem, one can conclude that the *base-matching* model and the *fc-matching* model themselves are equivalent.

Q.E.D. □

Compared with the assignment problem, the base model is more flexible. It is not necessary to figure out whether $|I'| \leq |J'|$ or $|J'| \leq |I'|$ before writing out the model constraints. Instead, one can consistently use the $\leq$- form of cardinality constraints.

In the literature, the authors often did not provide a specific value of M. A corollary of Observation 1 is that the specific value of M in the *fc-matching* problem does not matter, as long as it is greater than the cutoff distance c. The condition $M > c$ is necessary to ensure that exactly $|I'|$ assignments will be made. Otherwise, in the *fc-matching* problem, there will not be sufficient incentive to have any non-zero flows; and in the *base-matching* model, the conversion of constraint (40) to constraint (43) cannot happen.

3.1.2. A m:1 (1:n) Version of the Base Model

For exposition, a second base model for making m:1 and 1:n matches is also presented. Using the common notation in Section 2.2.1, the *base-bimatching* model can be defined in MILP as follows:

$$maximize\ Z = \sum_{(i,j) \in F} b_{ij} y_{ij} + \sum_{(i,j) \in G} b\prime_{ij} z_{ij} \tag{45}$$

Subject to:

$$\sum_{(i,j) \in F} y_{ij} \leq 1, \text{ for each } i \in I \tag{46}$$

$$\sum_{(i,j) \in G} z_{ij} \leq 1, \text{ for each } j \in J \tag{47}$$

The *base-bimatching* model is basically a fusion of two sub-models that maximize the total benefit of making forward assignments y_{ij} and backward assignments z_{ij}, respectively. The two sub-models are merged such that the objective function (45) maximizes benefits for both assignments simultaneously. Both the forward and backward assignments are partial assignments. Therefore, in (46), one can only require that the source object $i \in I$ should be assigned no more than once, and similar cardinality constraints cannot be imposed on the target $j \in J$. Likewise, for the backward assignments z_{ij}, one can impose cardinality constraints in only one direction.

The basic *base-bimatching* model is rather weak, as the model is free to make opposite assignments y_{ij}, z_{ij} without regard to each other. If a target $j \in J$ has been assigned to (with y_{ij}), nothing stops it from being assigned again as a source (via z_{ij}). Nonetheless, it is a common basis on which the many-to-one conflation models in the literature are formulated.

3.2. Link to the Assignment Problems

Next, this section presents an analysis of the links between existing optimized conflation models and the base models described above. From the proof of Observation 1, it is clear that the *base-matching* model is equivalent to a restricted assignment problem. This means that the assignment problem is equivalent to a *base-matching* model with appropriate modifications. In particular, if one sets the cutoff distance c to infinity, then the *base-matching* problem reduces to an assignment problem (1), (2) and (4), assuming $|I| \leq |J|$. If $|I| \geq |J|$, the roles of I and J must be swapped to obtain an assignment problem.

3.3. Link to the Network Flow Models

First of all, it should be noted that the general network flow problem can already be expressed as an MILP problem as follows:

$$\text{minimize} \sum_{e \in E} c_e f_e \tag{48}$$

Subject to

$$\sum_{e \in I_n} f_e - \sum_{e \in O_n} f_e = 0 \text{ for each } n \in N \tag{49}$$

$$l_e \leq f_e \leq u_e \text{ for each } n \in N \tag{50}$$

$$f_e = \text{ amount of flow in } e \in E$$

where f_e is the amount of flow along $e \in E$, N is the node set, E is the edge set, I_n, O_n are the sets of incoming and outgoing edges for the node n, respectively. While the objective (48) minimizes the total flow cost c_e, the only constraint (49) maintains flow conservation at each node.

However, the above formulation is not very useful, because it does not specify the specific network structure (in N and E) that is necessary for conflation. Any network flow based model could be expressed using the same MILP formulation in (48)–(50). A more useful connection to MILP models is the equivalence between the *base-matching* and *fc-matching* problems in Observation 1. By that proof, the *base-matching* problem *is the same* as the *fc-matching* problem.

3.3.1. Connection between fc-Matching and Base-Matching

As is proven in Section 3.1.1., the *fc-matching* is functionally the same as the *base-matching* model. They are essentially the same model expressed in two different formats (network flow and MILP, respectively).

3.3.2. Connection between fc-Bimatching and Base-Bimatching

$$maximize\ Z = \sum_{(i,j) \in F} b_{ij} y_{ij} + \sum_{(i,j) \in G} b\prime_{ij} z_{ij} - \beta \cdot \left(\sum_{(i,j) \in F} e_j + \sum_{(i,j) \in G} e\prime_i \right) \tag{51}$$

Subject to: (46), (47), and multi-assignment definitions

In the above, "multi-assignment definition" is a set of constraints that can correctly define two new variables s_j, e_j where

$s_j = 1$ if at least one source object i is assigned to j, or zero otherwise (for each $j \in J$).

$e_j =$ the amount of assignment to j that exceeds 1

Then, in the objective function (51), penalty terms were added to discourage excess assignments (e_j), as with the *fc-bimatching* problem. The decision variables for the primary assignment s_j and excess assignments e_j are characterized by two new constraints:

$$\sum_{(i,j) \in F} y_{ij} = s_j + e_j, \text{ for each } j \in J \tag{52}$$

$$M \times (1 - s_j) + e_j \leq M, \text{ for each } j \in J \tag{53}$$

Constraint (52) states that, for a target feature $j \in J$, the sum of the primary and excess assignments is the same as the total number of assignments. Constraint (53) uses the "big M" method and maintains that if there are non-zero excess assignments, the primary assignment variable s_j must be one.

Symmetrically, the primary assignment $s\prime_i$ and excess assignments ($e\prime_i$) can be defined for backward assignments.

$$\sum_{(i,j)\in F} z_{ij} = s\prime_i + e\prime_i, \text{ for each } i \in I \tag{54}$$

$$M \times (1 - s\prime_i) + e\prime_i \leq M, \text{ for each } i \in I \tag{55}$$

With the new variable for excess assignments and the associated penalty terms in the objective function, the *fc-bimatching* problem can be expressed as a variant of the *base-bimatching* problem.

3.4. Link to the Unified Bidirectional Matching u-Bimatching Problem

The unified bidirectional matching problem can be viewed as a merge of the two base models *base-matching*, *base-bimatching* plus additional constraints to maintain consistency between the full and directional assignments x_{ij}, y_{ij} and z_{ij}. The objective function (9) is a weighted sum of the objectives of the two base models. The additional constraints are the forward-backward compatibility constraints (16) and (17) plus constraints (14) and (15) as defined in Section 2.2.4.

3.5. Link to the Edge Connectivity Based Matching (ec-Matching) Problem

The *ec-matching* problem is a direct extension of the *base-matching* problem with added adjacency constraints (21) and (22) defined in Section 2.2.5.

The *ec-bimatching* problem is a direct extension of the *base-bimatching* problem with similar adjacency constraints (26) and (27) defined in Section 2.2.6.

3.6. Link to the Edge-Node Matching (en-Matching) Problem

The *en-matching* problem is a combination of two versions of the *base-matching* problem plus incidence constraints (33), (34), (35), and (36) as defined in Section 2.2.7. The two versions of *base-matching* are for matching the edges and the junctions of two networks, respectively.

As a summary of this section, Table 1 presents the main points about the relation between the existing models and the two base models. Column 1 presents the name of an existing model. Column 2 presents the base model based on which the said model can be built. Column 3 tabulates the additional constraints that are necessary during the model transform. Lastly, column 4 presents additional modifications in the objective function or the overall structure of the model.

4. Experiments

4.1. Experimental Settings

This section presents experiments that verify the theoretical links between existing conflation models and the properties of the two common base models. As detailed in the previous section, the existing MILP based conflation models in the literature are linked to the two base models in a straight-forward way: by adding additional constraints and merging objective function terms. Anyone interested in implementing one of these models can implement the base-matching models first, and then incrementally add the constraints and the objective function terms. What remains to be verified is the connection between the network-flow based *fc-matching* models and the base models. While it was argued that the *fc-matching* and *base-matching* models are equivalent, and their bi-matching versions are compatible, this needs to be verified computationally, since they are based on different

optimization algorithms. After all, the *fc-matching* models are based on network-flow solvers (such as the lemon C++ library), whereas the *base-matching* models are formulated and solved in MILP. Some of the implementation issues encountered during this cross-comparison are also discussed.

In order to verify the equivalence and relationships between the models, this article uses the recall and precision metrics and check whether the involved models generate the same outcome. The recall rate is used to measure the ability of the conflation algorithm to capture true matches. Given the number of true matches (*TM*), false matches (*FM*), and False Unmatches (*FU*), it is defined as:

$$\text{Recall} = \frac{TM}{TM + FU}$$

The precision rate is used to measure the algorithm's ability to be selective and include as few false matches as possible. It is defined as:

$$\text{Precision} = \frac{TM}{TM + FM}$$

As mentioned in the previous section, all the main network flow and MILP based models were implemented by the authors either by adapting existing code or developing new code when necessary. The implementation of the network flow based *fc-bimatching* models is based on the code in [11]. The remaining models were MILP-based and implemented using the Relational Linear Programming (RELP) package [32]. All experiments were conducted on a machine with an Intel i5-12400F CPU and 64 Gigabytes of system memory. For the test data, the same Santa Barbara County dataset with six test sites was used as in [11], as shown in Figure 3. Each site contains road networks from Open Street Map and TIGER, respectively.

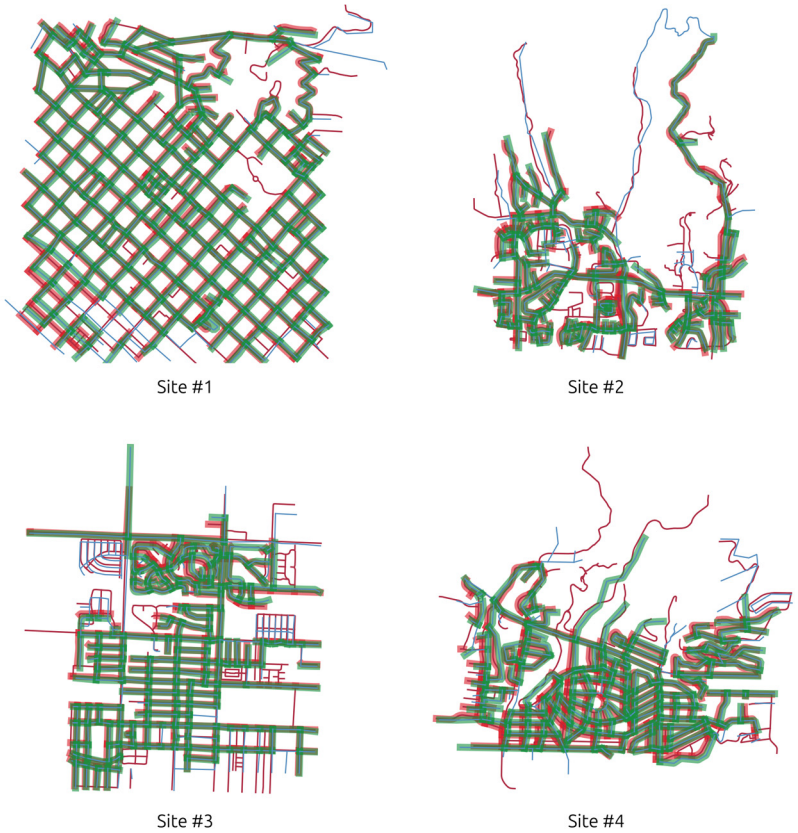

Figure 3. *Cont.*

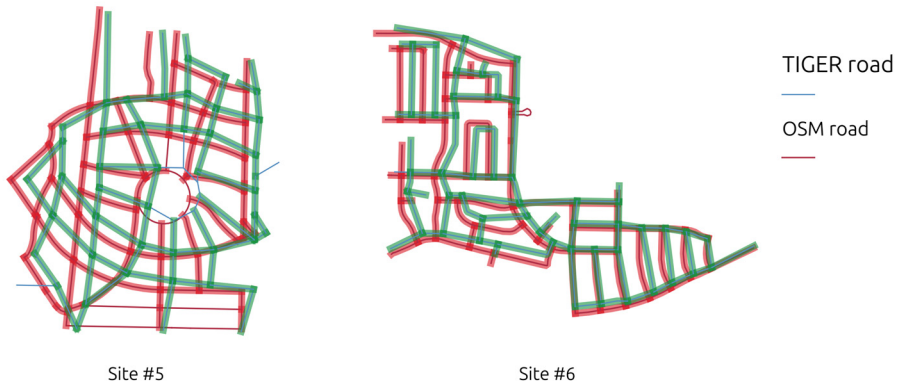

Figure 3. Six road datasets in Santa Barbara County, CA: Open Street Map vs. TIGER.

4.2. Properties of the Base Models

The first experiment is to test whether the fixed cost F in the base model (i.e., the *fc-matching* model) has any impact on the model outcome. This is important to test, as no recommendation has been given in the literature as to how the parameter should be chosen. A range of fixed cost values were tested ranging from 200 m to 1000 m with the original network flow-based *fc-matching* model. For each value of F, a range of cutoff distances from 40 m to 200 m at 20 m intervals were tested. Results for only the largest and the smallest sites (sites 1 and 5) are presented to save space, as the patterns are the same in all other sites.

Figure 4 presents the recall and precision curves (versus cutoff distance) at each F value. Firstly, one can observe that the performance curves (recall or precision) for different F values coincide. This verifies the claim made in the previous section that the fixed cost has no impact on the model outcome, as long as it is greater than the cutoff value c. Secondly, one can observe the general trend of recall rates increasing with the increase in the cutoff distance value c. The precision rate generally decreases after the cutoff distance reaches a certain level. When the cutoff distance is extremely small (e.g., Site 5 at $c = 40$), it may be the case that there are so few match candidates to choose from that the *fc-matching* model makes incorrect choices.

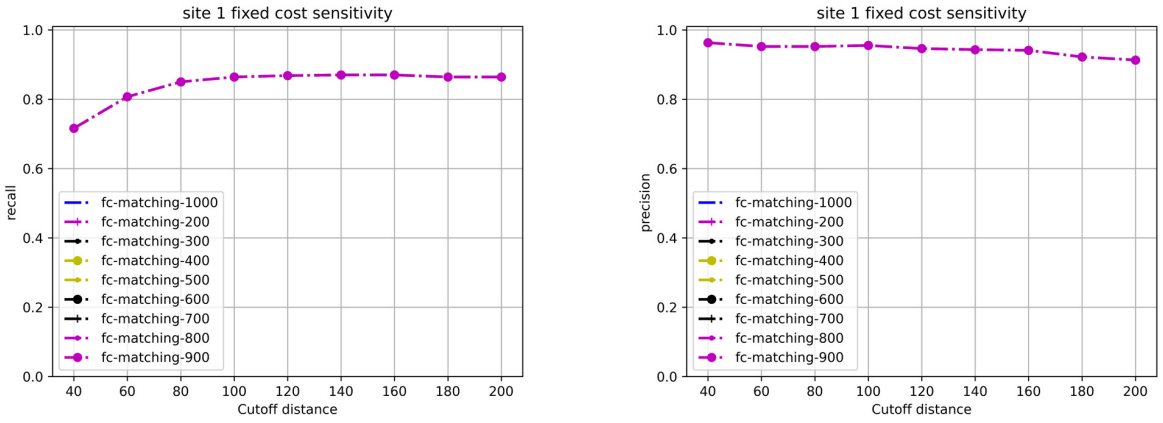

Figure 4. *Cont.*

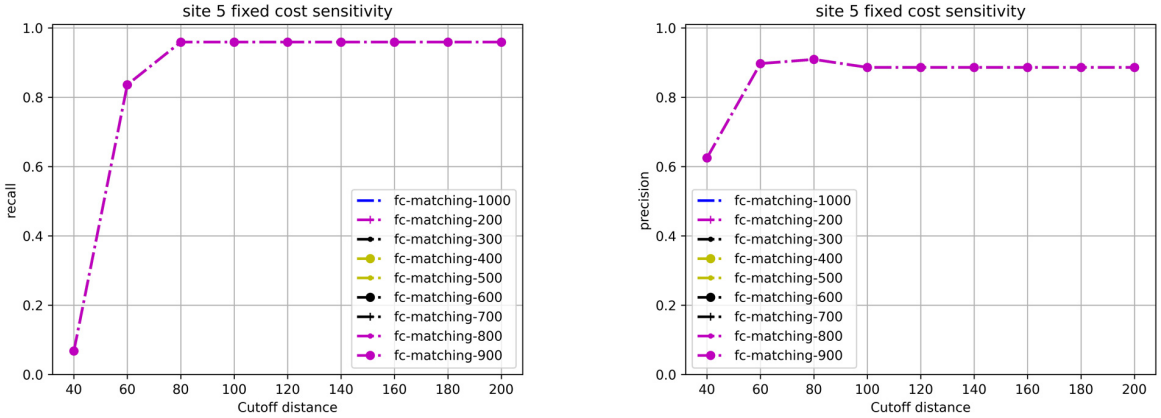

Figure 4. Sensitivity of the *fc-matching* model to the fixed cost (Sites 1 and 5).

4.3. Equivalences and Relations between the fc-Matching Problems and the Base-Matching Problems

The next step is to verify the equivalence between the MILP-based *base-matching* problem and the *fc-matching* problem, as well as the relations between their bi-matching versions. Figures 5 and 6 present the recall and precision rates for the *fc-matching*, *fc-bimatching*, *base-matching*, *base-bimatching*, as well as the MILP version of the *fc-bimatching* problem in (51) through (55), called *milp-fc-bimatching*. The first two models (*fc-matching* and *fc-bimatching*) are network-flow based, whereas the other three models are MILP-based. Firstly, one can observe that in all figures, the MILP-based *base-matching* and the network-flow based *fc-matching* models have almost identical recall and precision rates at each cutoff distance. This fact experimentally verifies our hypothesis that the two models are logically identical.

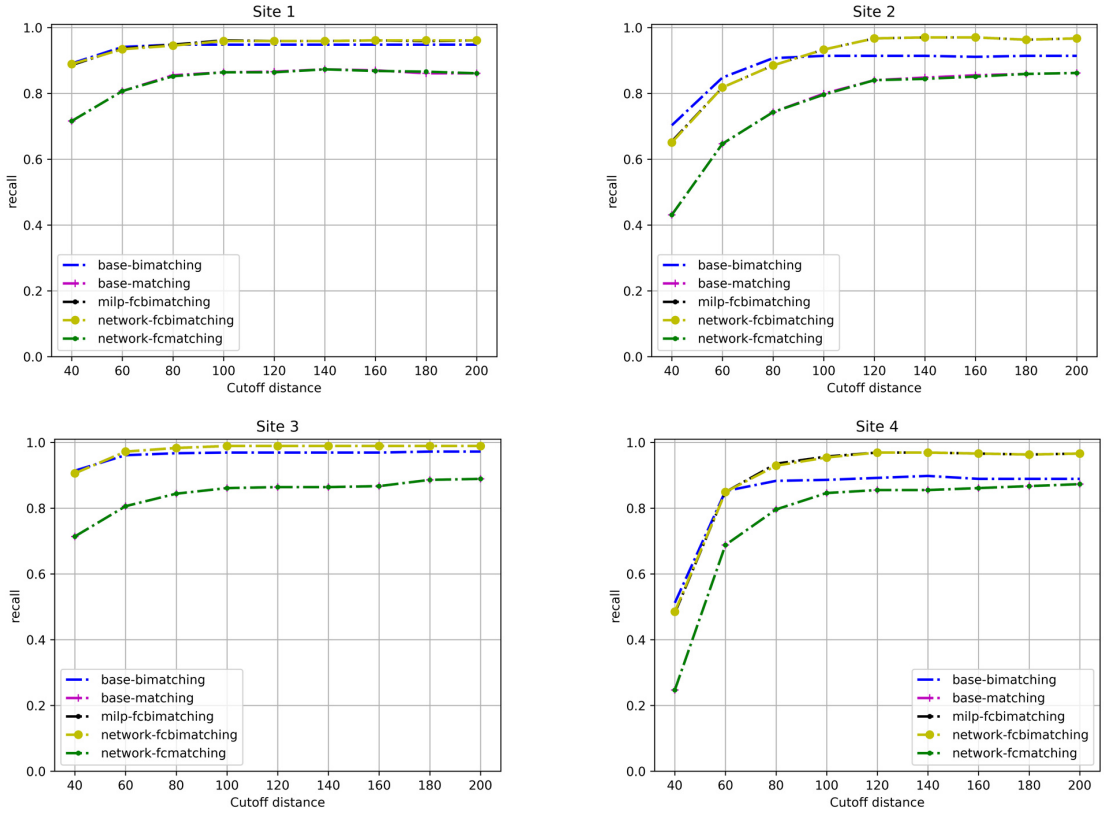

Figure 5. *Cont.*

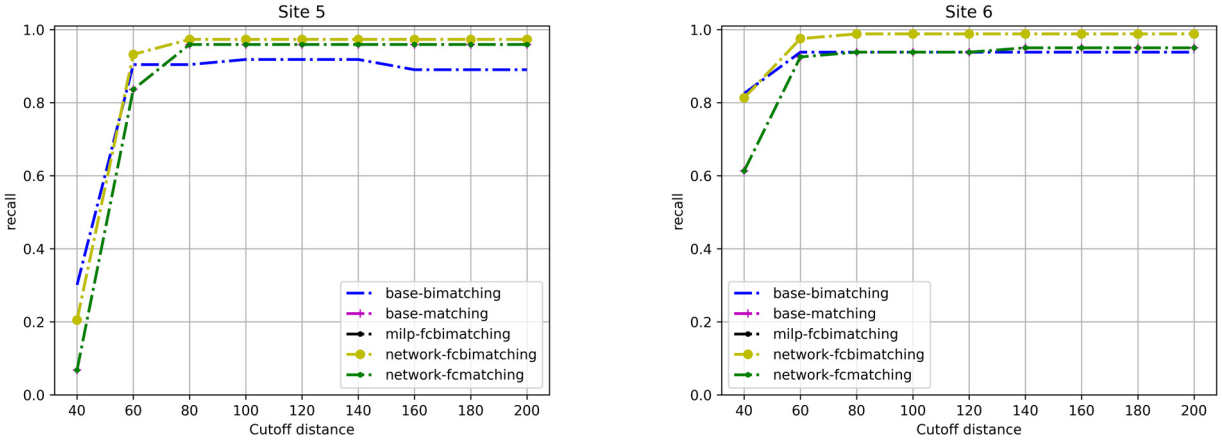

Figure 5. The recall for the *fc-matching* problems and the *base-matching* problems.

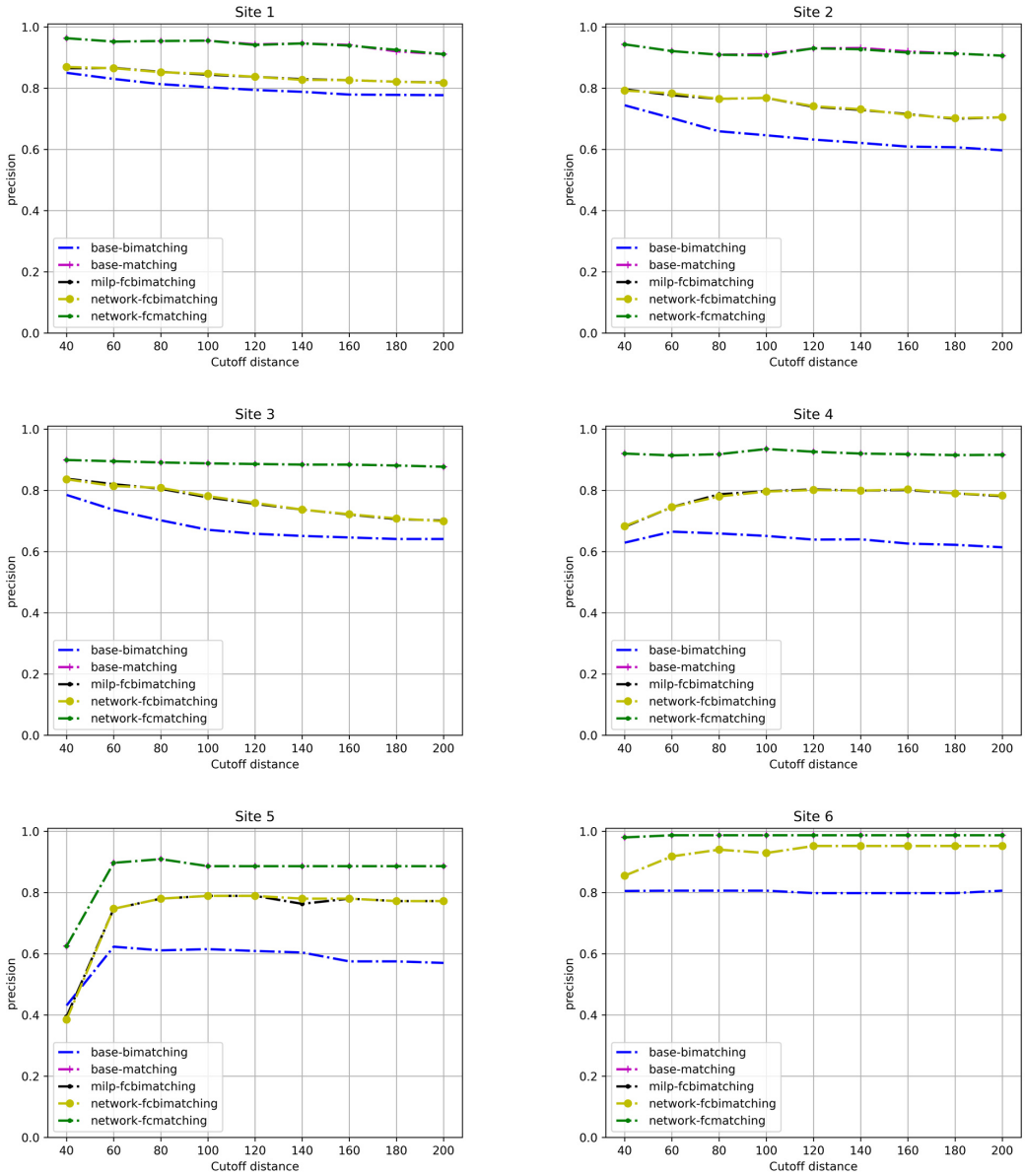

Figure 6. The precision for the *fc-matching* problems and the *base-matching* problems.

For the bi-matching models, one can observe likewise that the MILP-based *milp-fcbimatching* model and the original network based network flow model have identical recall and precision rates at all six test sites, except for small differences in site 5 at cutoff distances 40 m and 100 m, respectively. Upon closer investigation, the difference arose from two related implementation issues: integer round-off and distance ties. More specifically, the networkflow-based models in [11,15] are based on the Lemon C++ library for solving network problems. As with many code bases and algorithms of its kind, the specific minimum cost circulation algorithm from that library assumes that the flow costs on network edges are integer valued. Therefore, the network flow models in [11,15] effectively round off distances to integer values beforehand. To cope with this issue, the distances are also rounded off in our MILP based models, but a new problem emerged: ties between integer valued distances. In some sites such as site 5, the directed Hausdorff distance to two different target features can be the same. Since the different conflation algorithms make arbitrary choices to break the tie, the prescribed matches may be different, therefore, about these features.

Practically, the experiment above illustrates some nuances between the solvers and conflation models. Even though the two models are logically equivalent, the model results may vary due to limitations in some of the solvers. In this case, the network flow-based solver requires integer valued input, whereas the MILP based solver does not suffer from this limitation. On the other hand, specialized network-flow based solvers are theoretically faster than the more general MILP based solvers for the same problem at hand.

Figures 5 and 6 also demonstrate the differences between the two base models as well as the effect of extending the base models. Generally, the plotted performance curves form two groups based on the cardinality of matching, with the 1:1 models (*fc-matching* and *base-matching*) being one group and the m:1 (1:n) models being the other. In terms of the model outcome, the 1:1 models are identical, while the *base-bimatching* and the *fc-bimatching* problems (MILP or network-based) are close but have some differences. Within the m:1 group, the recall rates for the *base-bimatching* problem are consistently higher than those for the *fc-bimatching* problem. The only exception is at site 5 when the cutoff distance is very small (40 m), in which case the candidate match set may have been overly cut down. Symmetrically, the precision rates for the *base-bimatching* problem are consistently lower than those for the *fc-bimatching* problem. This difference reveals the effect of the added constraints (52) through (55) on the base bi-matching model. As intended, they identify excess assignments (multi-assignments), which are then penalized in the objective. As expected, this increases the precision as the unreliable multi-assignments are discouraged, but reduces recall (also because of the reduced multi-assignments). From a perspective, the *base-bimatching* problem is the most unrestrictive model among bi-matching models and is probably too unrestrictive. For example, at site 5, its precision is almost 20% lower than that of the *fc-bimatching* problem (at a 100 m cutoff). On the other hand, this comparison also shows the effectiveness of the various add-on features to the base models.

5. Conclusions and Future Directions

Optimized conflation is a method for GIS data conflation, which is aimed at matching two GIS datasets by systematically minimizing the total discrepancy. First conceptualized in the 1980s, optimized conflation has seen many different formulations of the conflation problem, starting from the map assignment problem to the more recent network-flow based and newer models based on MILP. This paper demonstrates that all the present optimized conflation models are intrinsically linked. In particular, two base models are presented (for the 1:1 and m:1/1:n matching cases), which serve as a common ground for all existing models. One base model (the *base-matching* problem) is the network-flow based *fc-matching* problem reformulated in MILP. The other base-model (the *base-bimatching* problem) is an m:1 version of the base model. By means of the base conflation models, this article showed that existing models can be viewed as variants of the base model(s) with either added constraints/objective terms, or modified problem parameters.

Since the 1:1 base model (*fc-matching*) was originally networkflow-based, it was necessary to demonstrate that it can be succinctly expressed in MILP. The Method section demonstrates that indeed, *fc-matching* can be reformulated in MILP, resulting in a model structure that happens to be the common ground for the existing 1:1 models. It is also experimentally verified that the network-flow-based and MILP-based versions of the base model generated the same outcome. For the m:1 case, no existing model could serve as a common base model. Instead, a bi-matching version of the base model was presented (called *base-bimatching*), of which the existing bi-matching models can be viewed as extension models. The properties of the base models were then discussed and compared with each other.

The contributions of this article are three fold. Firstly, the identification of the common base model helps to understand the multitude of optimized conflation models in the literature. In light of the base models, one can clearly see what a specific conflation model adds to the baseline model. Secondly, the common base models ease the implementation of optimized conflation models by promoting a modular way of model development. One can implement a base model and then add additional features incrementally. Thirdly, practical issues were discussed including the choice between network-flow and MILP based solvers, as well as the choice of basic parameters, such as the fixed cost.

Author Contributions: Conceptualization, Ting L. Lei; Methodology, Zhen Lei and Ting L. Lei; Software, Zhen Lei, Zhangshun Yuan and Ting L. Lei; Validation, Zhen Lei and Ting L. Lei; Formal analysis, Zhen Lei and Ting L. Lei; Investigation, Zhen Lei and Ting L. Lei; Resources, Ting L. Lei; Data curation, Zhen Lei and Ting L. Lei; Writing—original draft, Zhen Lei and Ting L. Lei; Writing—review & editing, Zhen Lei and Ting L. Lei; Visualization, Zhen Lei, Zhangshun Yuan and Ting L. Lei; Supervision, Zhen Lei and Ting L. Lei; Project administration, Ting L. Lei; Funding acquisition, Ting L. Lei. All authors have read and agreed to the published version of the manuscript.

Funding: This research was partly supported by Natural Science Foundation, Grant number BCS-2215155. This research was partly supported by National Natural Science Foundation of China (NSFC) Grant number 41971334.

Data Availability Statement: The data that support the findings of this study are available upon reasonable request.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Ruiz, J.J.; Ariza, F.J.; Ureña, M.A.; Blázquez, E.B. Digital map conflation: A review of the process and a proposal for classification. *Int. J. Geogr. Inf. Sci.* **2011**, *25*, 1439–1466. [CrossRef]
2. Xavier, E.M.A.; Ariza-López, F.J.; Ureña-Cámara, M.A. A survey of measures and methods for matching geospatial vector datasets. *ACM Comput. Surv.* **2016**, *49*, 1–34. [CrossRef]
3. Saalfeld, A. A fast rubber-sheeting transformation using simplicial coordinates. *Am. Cartogr.* **1985**, *12*, 169–173. [CrossRef]
4. Saalfeld, A. Conflation automated map compilation. *Int. J. Geogr. Inf. Syst.* **1988**, *2*, 217–228. [CrossRef]
5. Brown, J.N.; Rao, A.L.; Baran, J. Automated GIS conflation: Coverage update problems and solutions. In Proceedings of the 1995 Geographic Information Systems for Transportation (GIS-T) Symposium, Washington, WA, USA, 2–5 April 1995.
6. Masuyama, A. Methods for detecting apparent differences between spatial tessellations at different time points. *Int. J. Geogr. Inf. Science* **2006**, *20*, 633–648. [CrossRef]
7. McKenzie, G.; Janowicz, K.; Adams, B. A weighted multi-attribute method for matching user-generated points of interest. *Cartogr. Geogr. Inf. Sci.* **2014**, *41*, 125–137. [CrossRef]
8. Li, L.; Goodchild, M.F. Automatically and accurately matching objects in geospatial datasets. *Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci. ISPRS Arch.* **2010**, *38*, 98–103.
9. Li, L.; Goodchild, M.F. An optimisation model for linear feature matching in geographical data conflation. *Int. J. Image Data Fusion* **2011**, *2*, 309–328. [CrossRef]
10. Tong, X.; Liang, D.; Jin, Y. A linear road object matching method for conflation based on optimization and logistic regression. *Int. J. Geogr. Inf. Sci.* **2014**, *28*, 824–846. [CrossRef]
11. Lei, T.L. Geospatial data conflation: A formal approach based on optimization and relational databases. *Int. J. Geogr. Inf. Sci.* **2020**, *34*, 2296–2334. [CrossRef]
12. Wu, H.; Xu, S.; Huang, S.; Wang, J.; Yang, X.; Liu, C.; Zhang, Y. Optimal road matching by relaxation to min-cost network flow. *Int. J. Appl. Earth Obs. Geoinf.* **2022**, *114*, 103057. [CrossRef]

13. Rosen, B.; Saalfeld, A. Match Criteria for Automatic Alignment. In Proceedings of the 7th International Symposium on Computer-Assisted Cartography, Washington, WA, USA, 11–14 March 1985; pp. 1–20.
14. Hillier, F.S.; Lieberman, G.J. *Introduction to Operations Research*, 8th ed.; McGraw-Hill: New York, NY, USA, 2005; p. 1088.
15. Lei, T.; Lei, Z. Optimal spatial data matching for conflation: A network flow-based approach. *Trans. GIS* **2019**, *23*, 1152–1176. [CrossRef]
16. Lei, T.L.; Wang, R. Conflating linear features using turning function distance: A new orientation-sensitive similarity measure. *Trans. GIS* **2021**, *25*, 1249–1276. [CrossRef]
17. Ai, T.; Cheng, X.; Liu, P.; Yang, M. A shape analysis and template matching of building features by the fourier transform method. *Comput. Environ. Urban Syst.* **2013**, *41*, 219–233. [CrossRef]
18. Beeri, C.; Kanza, Y.; Safra, E.; Sagiv, Y. Object fusion in geographic information systems. *Proc. Thirtieth Int. Conf. Very Large Data Bases* **2004**, *30*, 816–827.
19. Corral, A.; Manolopoulos, Y.; Theodoridis, Y.; Vassilakopoulos, M. Algorithms for processing k-closest-pair queries in spatial databases. *Data Knowl. Eng.* **2004**, *49*, 67–104. [CrossRef]
20. Tong, X.; Shi, W.; Deng, S. A probability-based multi-measure feature matching method in map conflation. *Int. J. Remote Sens.* **2009**, *30*, 5453–5472. [CrossRef]
21. Lei, T.L.; Lei, Z. Harmonizing full and partial matching in geospatial conflation: A unified optimization model. *ISPRS Int. J. Geo-Inf.* **2022**, *11*, 375. [CrossRef]
22. Lei, T.L.; Lei, Z. Linear feature conflation: An optimization-based matching model with connectivity constraints. *Trans. GIS* **2023**, *27*, 1205–1227. [CrossRef]
23. Lei, Z.; Lei, T.L. Towards topological geospatial conflation: An optimized node-arc conflation model for road networks. *ISPRS Int. J. Geo-Inf.* **2024**, *13*, 15. [CrossRef]
24. Yang, B.; Zhang, Y.; Luan, X. A probabilistic relaxation approach for matching road networks. *Int. J. Geogr. Inf. Sci.* **2013**, *27*, 319–338. [CrossRef]
25. Zuo, Z.; Yang, L.; An, X.; Zhen, W.; Qian, H.; Dai, S. A hierarchical matching method for vectorial road networks using delaunay triangulation. *ISPRS Int. J. Geo-Inf.* **2020**, *9*, 509. [CrossRef]
26. Guo, Q.; Xu, X.; Wang, Y.; Liu, J. Combined matching approach of road networks under different scales considering constraints of cartographic generalization. *IEEE Access* **2020**, *8*, 944–956. [CrossRef]
27. Wang, Y.; Yan, H.; Li, P.; Lu, X. A multiscale road matching method based on hierarchical road meshes. *Earth Sci. Inform.* **2024**, *17*, 1765–1778. [CrossRef]
28. Ali, A.B.; Harvey, F.; Vauglin, F. Geometric Matching of Areas, Comparison Measures and Association Links. In Proceedings of the 8th International Symposium on Spatial Data Handling, Vancouver, BC, Canada, 11–15 July 1998; pp. 557–568.
29. Fan, H.; Zipf, A.; Fu, Q.; Neis, P. Quality assessment for building footprints data on OpenStreetMap. *Int. J. Geogr. Inf. Sci.* **2014**, *28*, 700–719. [CrossRef]
30. Song, W.; Keller, J.M.; Haithcoat, T.L.; Davis, C.H. Relaxation-based point feature matching for vector map conflation. *Trans. GIS* **2011**, *15*, 43–60. [CrossRef]
31. Fu, Z.; Yang, Y.; Gao, X.; Zhao, X.; Lu, Y.; Chen, S. Road networks matching using multiple logistic regression. *Geomat. Inf. Sci. Wuhan Univ.* **2016**, *41*, 171–177. [CrossRef]
32. Lei, T.L. Integrating GIS and location modeling: A relational approach. *Trans. GIS* **2021**, *25*, 1693–1715. [CrossRef]