AN ITERATIVE APPROACH FOR MATCHING MULTIPLE REPRESENTATIONS OF STREET DATA

S. Volz

University of Stuttgart, Institute for Photogrammetry, 70174 Stuttgart, Germany –
steffen.volz@ifp.uni-stuttgart.de

Commission II, WG II/3; WG II/6

KEY WORDS: Multiple Representations, Spatial Databases, Vector Data Matching

ABSTRACT:

In spatial data integration the most difficult problems arise due to multiple, inconsistent representations of one and the same real world object in different geospatial databases. One of the biggest challenges regarding the integration of multiple representations is the identification of corresponding objects within diverse source data sets. This process is generally referred to as spatial data matching. Some sophisticated approaches have been presented to cope with the task but still methods are needed to optimize the procedure. In this work, it is intended to achieve such an optimization by applying an iterative approach for the matching of street data form disparate sources, namely the Geographic Data Files (GDF) format and the German Authoritative Topographic Cartographic Information System database (ATKIS). After reducing the global geometric deviation of the linear source data sets by a rubber sheeting transformation, the street objects are topologically split and additional nodes are introduced, respectively, in order to enable the detection of a maximum number of 1:1 matches. Then, the matching process starts by identifying seed nodes in the source data sets which show a high likelihood of correspondence. With the seed nodes as starting points, a combined edge and node matching algorithm detects 1:1 correspondences. In case no 1:1 match could be found, an enhanced edge matching approach being able to recognize 1:2 matches is triggered. The whole process is performed in multiple iterations and it is repeated applying relaxed constraints. The results of the matching are stored in explicit relations expressing the degree of inconsistency of multiple representations.

1. INTRODUCTION

Spatial data infrastructures are evolving on the global (GSDI 05, Nexus 05), on the national (ADV 04, ANZLIC 05, FGDC 05) and also on regional or city levels. It is their common goal to unify existing data sets within one platform and to provide an integrated view on the underlying data. However, this common goal also involves a common problem: the existing data sets which have been acquired by different institutions according to different conceptual schemas and data models, in different formats and scales, with different accuracies or at different dates, etc. are highly inconsistent. Basically, two types of inconsistencies can occur: First, different object types could not fit together: imagine a small scale building data set which is overlaid with a large scale street data set, leading to invalid intersections between street and building objects. Second, the same real world object could be represented in two different spatial databases, leading to possible inconsistencies in case that attributes of these representations are contradictory. For example, the geometries of streets which have been captured by different companies or institutions in their individual databases will never be exactly the same. This raises the question which of the representations is the best approximation of the real world. Of course, that is depending on the application context, too.

This paper deals with the second inconsistency issue, the problem of multiple representations of the same real world phenomenon. To be able to achieve an integration of inconsistent representations available in existing data sets within a spatial data infrastructure, first of all corresponding instances have to be identified. This process is generally referred to as spatial data matching. It is highly problematic since situations can occur which even cannot be resolved by human operators.

In our approach, we try to find an optimized solution for the matching of street data sets which have been acquired according to two different conceptual schemas, namely the Geographic Data Files standard (GDF) and the German Authoritative Topographic Cartographic Information System database (ATKIS). Both GDF and ATKIS capture street objects as linear features at an approximate scale of 1:25000. The approach developed here encompasses several steps. In a first phase, the data are prepared to allow for an optimized matching process. Then, the matching itself is performed in several iterations to achieve the final result which is stored as explicit relations between corresponding features. The result set also includes those features for which no matching candidates could be found.

The remainder of this paper is organized as follows: in section 2, related work is discussed. Section 3 presents the investigated data and section 4 gives an overview of the data pre-processing steps. Section 5 contains a detailed explanation of the proposed matching process and in section 6 results and drawbacks of the approach are outlined. Finally, section 7 concludes the paper and gives an outlook on future issues.

2. RELATED WORK

Identifying corresponding objects in different data sets or data matching, respectively, is not unique to geospatial databases. Also in alphanumeric data like relational tables or semi-structured data (like XML documents) it is necessary to find correspondences between similar representations to perform data integration, for example in the area of genome databases. In the database domain, this process is rather referred to as data cleansing, record linkage or duplicate detection. An example for the detection of corresponding elements in nested XML documents can be found in (Weis and Naumann 04). The authors apply a threshold-based similarity function which relies on the edit distance measure (that basically calculates the number of changes needed to transform a source string into a target string) in order to identify corresponding string objects. To reduce the number of expensive edit distance computations different filter methods have been implemented. Thus, corresponding instances and structures can be identified efficiently.

In alphanumeric databases, usually all records within different databases have to be compared with each other. However, in spatial data matching, the number of pairwise comparisons of duplicate objects or multiple representations, respectively, can be significantly reduced since a simple fact can be exploited: two objects representing the same real world phenomenon have to share at least approximately the same location on earth, and thus the search window and the number of possible matching candidates, respectively can be minimized. However, the problem still remains very difficult.

Multiple representations of spatial data are mainly resulting from the fact that different geospatial communities are interpreting the real world from their individual perspectives (Bishr et al. 99). Thus, they create their own conceptual schemas based on which they acquire the data. However, there are also other reasons for the occurrence of multiple representations, like the fact that the same real world object can be captured at different dates. Multiple representations can vary in many different ways and to different degrees: they can have different geometries or geometry types, different scales, different semantics and different relational properties like e.g. different topologic relations, etc.

Generally, a manual approach for the identification of corresponding geospatial objects is considered to be most promising. However, due to the huge amounts of spatial data available and due to their high update rates, this approach is not applicable. Thus, automatic matching techniques have to be developed. In the following sections, some existing methods for the matching of multiple representations are presented and differentiated on the basis of the geometric types of the objects to be matched. Just like the research presented in this paper, these approaches mainly focus on data of similar scale. There are other projects dealing with multi-scale issues like (Jones et al. 96, Dunkars 03), but they shall not be discussed here in detail.

2.1 Point-based methods

Point-based methods generally consider the proximity of the points to be matched and also investigate the properties of the incident features in case there are such. One algorithm which can be assigned to this category is based on the concepts that have been developed within the EVIDENCE project (Pandazis 99). It has been implemented by (Bofinger 01). The algorithm is based on the idea of describing intersections of streets, i.e. nodes of a street network, by an explicitly defined code. The code consists of point coordinates, abbreviations and names of incident streets and the number of linked edges. For each intersection, such a code is created. By comparing the codes of features within different data sets and by assigning the intersections with the most similar codes to each other, references can be derived.

In (Beeri et al. 2005), location-based database join algorithms for point datasets are developed. They are also capable of matching more than two data sets at a time, either in a sequential or a simultaneous fashion. The performances of the algorithms are presented in terms of recall and precision.

2.2 Line-based methods

In (Walter 97, Walter and Fritsch 99) a fundamental, line-based matching approach for street network data of ATKIS and GDF has been presented. In a first step, the algorithm finds all potential correspondences of topologically connected line elements in two source data sets by performing a buffer operation. The matching candidates are stored in a list. This list is ambiguous and typically contains a large amount of *n:m* matching pairs. Then, unlikely matching pairs are identified and eliminated using relational parameters like topological information and feature-based parameters like line angles. The result is a smaller but still ambiguous list with potential matching pairs. These matching pairs are evaluated with a merit function in order to compute a unique combination of matching pairs which represents the solution of the matching task. This is a combinatorial problem which is solved with an A* algorithm (see Aho et al. 87).

The buffer algorithm of (Walter 97) has recently been adapted by several other authors. (Mantel & Lipeck 04) extend the algorithm to be able to apply it in a symmetric fashion for the matching of cartographic objects. In (Stigmar 05), the Java Conflation Suite developed by the Jump Project (JUMP 05), is extended by 3 different modules, one of which also uses the buffering approach to optimize matching procedures between route data derived from navigation systems and road data provided by national mapping agencies. Also, (Zhang et al. 05) apply the buffer algorithm while matching street networks. They developed a method to adjust the buffer parameters during the matching process to find an optimal solution.

2.3 Area-based methods

In (Kraft 95) corresponding areas in two datasets are used in order to minimize global geometric differences between these two datasets. All areas which have a distance less than a specific threshold are interpreted as potential matching pairs. All potential matching pairs are evaluated with a cost function which is calculated by different weighted parameters like size of area, centre of gravity, Kappa number or number of line segments. This leads again to an ambiguous list of matching pairs. This list is sorted by the costs, beginning with the matching pair with the lowest costs. The matching with the lowest costs is used for the final result and all remaining matching pairs, which contain one of these areas, are eliminated until the list is empty.

Further work on matching area objects was proposed by (van Wijngarden et al. 97) who matched building objects by calculating the percentage of overlap of the building geometries. In (von Gösseln and Sester 04) an approach to match water areas of ATKIS and geological/soil databases with cardinalities up to n:m was introduced. They apply several similarity measures like degree of overlap, area-to-perimeter ratio or the Hausdorff distance to detect corresponding objects.

2.4 Mixed Approaches

Mixed approaches combine point-based and line-based or area-based matching procedures. (Filin and Doytsher 00) adopt similar algorithms for the geometric matching of nodes as described in 2.1. They also present an advanced node matching approach that utilizes topological properties for identifying counterpart nodes. It performs a "walk" from one source node (e.g. A) in a data set *a* to its adjacent nodes (e.g. B and C) and then to their corresponding nodes in data set *b* (B' and C'). The match is considered valid if direct connections from B' and C' to target node A' in data set *b* exist. Thus, corresponding nodes which cannot be found by applying the proximity criterion (i.e. which are located outside the search window) still are detectable. The authors augment their node matching algorithm by a topology-based approach for the detection of corresponding edges. It tries to cope with the problem of fragmentation of linear features into segments. First, the shortest path between two nodes A and B in one data set is found and then the most similar path between the corresponding nodes A' and B' in the other data set is identified by calculating the relative area between the paths.

In (Xiong and Sperling 04), node, segment and edge matching algorithms are combined as well. Their purpose is to match linear road features extracted from aerial photographs and existing street network databases. During a first phase sets of corresponding node features are grouped into clusters or sub-networks, respectively. This concept is based on the notion that matches should rather be carried out on the basis of the perception of a whole spatial situation, and not only by looking at individual cases. In a further step, the clusters themselves are compared and their "closeness" is evaluated. Then, the method identifies a set of highly corresponding seed nodes within corresponding clusters by geometric and topological criteria. This step is supervised by a human operator who can insert undetected seed nodes and delete incorrect matches after the automatic procedure has been finished. Starting from the seed nodes, the edge matching is performed. Those candidates are accepted as edge matching pairs which show the highest degree of similarity with respect to angle and length difference criteria as well as distance approximations.

Another example of a mixed approach was implemented by (Kraut 03). First, it identifies junctions of different street networks with a high likelihood of correspondence by comparing strictly defined geometric, attributive and topologic parameters. The found pairs are defined as start points. They are used for an affine transformation to adjust the data sets. Basically, running in each direction from any start point and comparing each adjacent node and again their adjoining nodes, and so on, the whole network can be examined. If both the start and end nodes of two streets are identified as being coincident, also the edges are matched. The approach proposed in this paper adopts a similar principle.

3. INVESTIGATED DATA

In our approach, we investigated street data stemming from different conceptual schemas, namely ATKIS and GDF. The data sets have been captured in approximately the same scale (see Figure 1, showing a clipping of the test data). ATKIS and GDF are briefly introduced in the following sections.

3.1 ATKIS

ATKIS is a general topographic database that stores data of different topographic object categories like vegetation, settlements, traffic, etc. It is not targeted to a certain application domain but rather serves as an information basis on top of which application-dependent data can be added. The ATKIS data are being captured in different digital landscape models (DLM) with scales of 1:25000 (used in this research), 1:50000, 1:250000 and 1:1000 000. ATKIS is hierarchically organized and is an object-based system, but it does not support inheritance.

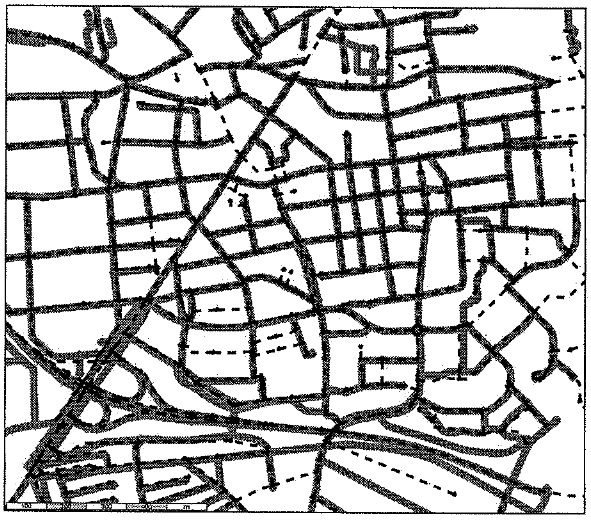


Figure 1. ATKIS (▬▬) and GDF (– – –) street data sets (situation after alignment has been performed, see section 4.1)

The topographic object categories (like 'traffic') are subdivided into so-called object groups (e.g. road traffic, air traffic, etc.) which contain the object classes (like Street, Way, etc.). The instances of the object classes are the objects. On the object level, the geometries of objects are stored and alphanumeric attributes, e.g. the order of a road or its width, are given.

3.2 GDF

The conceptual schema of GDF is focused on describing road networks for car navigation purposes and therefore contains all the information necessary to perform routing. The central element in GDF is the "feature" which corresponds to some kind of real world object like a railway, a street, etc. Real world objects represented by a point, line or area are called simple features, while complex features are composed of a group of simple features. Features contain geometric as well as thematic information and can be linked by relations. Every feature belongs to a specific feature class (like Road Element, Road, etc.), i.e. GDF can also be considered as a hierarchical, object-based system but just like in ATKIS inheritance is not supported. GDF data are captured by 2 different companies, namely NAVTEQ and TeleAtlas, with a positional accuracy of approximately ±3 metres.

4. DATA PRE-PROCESSING

In order to optimize the matching process, some data preparation processes had to be carried out on the source data sets: first, the global geometric deviation between them had to be reduced. Second, the linear street features within both data sets had to be split to achieve as many 1:1 matches as possible.

4.1 Reducing the Geometric Deviation

In order to reduce the global geometric deviation between data sets to be matched, they first have to be adjusted. Thus, the area in which corresponding features have to be looked for can be reduced and ambiguities can be minimized. This was done by a rubber sheeting transformation (see e.g. Cobb et al. 98).

For the alignment process, so-called warping nodes had to be found in both data sets which were showing a high degree of correspondence. All those nodes were considered as potential warping nodes which were located within a distance of 100 metres from each other and which had at least 4 incident edges showing approximately the same length and the same angles. In order to be accepted as warping nodes the criterion of mutual unambiguousness (see section 5.1) had to be fulfilled. Finally, warping vectors between the corresponding ATKIS and GDF warping nodes were created. After having identified all potential warping vectors, we compared them with respect to their length and orientation and removed strong outliers. On the basis of the resulting warping vectors set, the rubber sheeting transformation was performed.

4.2 Topological splitting

Basically, a representation can either be made up of one or of multiple individual objects. Thus, different cardinalities from 1:1 up to n:m can occur during matching procedures. Determining n:m matches, though, is much more difficult than identifying 1:1 relations since it results in a combinatorial problem as it was shown in the approach of (Walter and Fritsch 99). For this reason, we intended to split street features so that as many 1:1 matches as possible can be found.

The basic algorithm that solves this task is depicted in Figure 2. In Figure 2 (a) two ATKIS and two GDF edges and their start and end nodes are displayed. In this situation, a correct match is hard to be detected. The situation improves if more nodes are introduced and the edges are split, respectively. This process contains multiple steps (see Figure 2 (b)): first of all, object a_1 of the ATKIS map is buffered (a_1 buffer). Then, all GDF edges are determined that intersect this buffer and show approximately the same angle (± 10 degrees). This is only object g_2. All segments of a_1 buffer that intersect either the start or the end node of a_1 are taken as input segments and an intersection between them and g_2 is calculated. If an intersection could be detected *and* if within the search area around the intersection point no node of edge g_2 can be found, a new GDF node is introduced – and g_2 is split into g_2' and g_4. The same process is carried out for ATKIS edge a_2. However, in this case, no new node is created since the intersection points of the buffer segments are located within the search area around node g_{n3} (see black cross) and within the search area of the newly introduced node (actually it is located at the same position). The same algorithm is also performed for the GDF edges which leads to a new ATKIS node and, consequently, to the splitting of edge a_1 into the edges a_1' and a_3. If objects are split, the attributes of the existing objects are simply copied and transferred to the newly created objects. The resulting situation in Figure 2 (c) now shows two clear matches, namely $a_3 \leftrightarrow g_4$ and $a_2 \leftrightarrow g_2$'. The edges a_1' and g_1 cannot be assigned since their angle difference is too large.

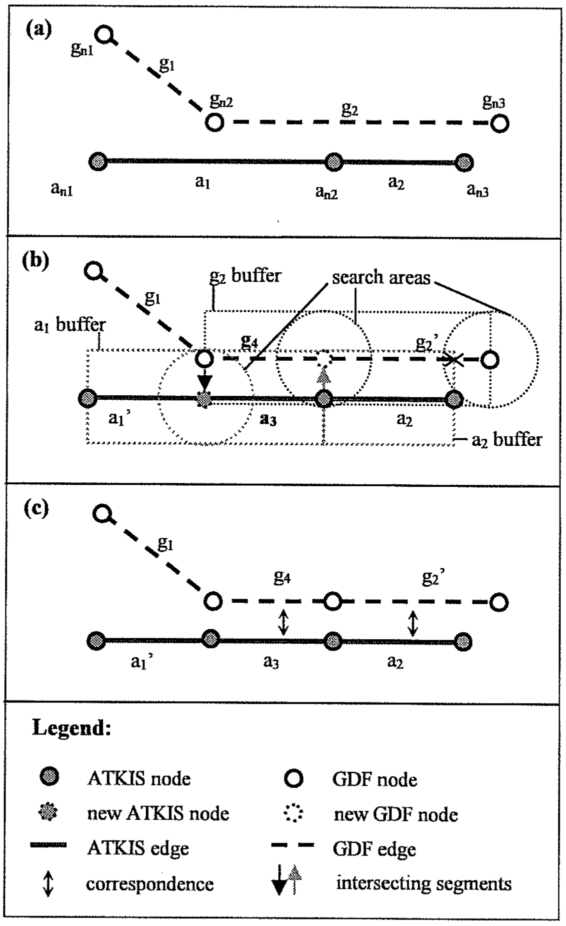


Figure 2. Splitting edges by transferring nodes from one representation to the other

5. THE ITERATIVE MATCHING APPROACH

The matching approach proposed in this paper consists of different steps. First, so-called seed nodes (according to the terminology of Xiong and Sperling 04) which show a high likelihood of correspondence are identified. Then, all edges emanating from the seed nodes and again their end nodes are compared and so on. Following this principle, the complete network can be examined by performing multiple iterations. The process is explained in detail in the following sections. Additionally, the way we store links between corresponding objects is outlined. Finally, we briefly describe how we handle cases which cannot be correctly detected by the automatic system.

The whole implementation of the approach has been carried out within the Java-based open source GIS environment JUMP (JUMP 05). None of the algorithms provided by the Java Conflation Suite or the Road Matcher Application which have been developed within the JUMP project were used during the pre-processing and matching steps except for the computation of some geometric similarity measures.

5.1 Finding seed nodes

The algorithm that detects the seed nodes is basically similar to the one that finds corresponding nodes for performing the rubber sheeting transformation. First, all nodes of the ATKIS data set are considered which have more than 2 incident edges. For all those nodes, corresponding GDF nodes which are located within a 30 metres distance and show exactly the same number of emanating edges with almost the same angles (a tolerance of ± 5 degrees is accepted) are selected as potential candidates. Then, the same process is carried out vice versa, i.e. all corresponding ATKIS nodes are being detected for the GDF nodes. In any case where only one possible counterpart could be found for both the ATKIS and the GDF node, the corresponding node objects are characterized as seed nodes. So the algorithm works bi-directional and thus meets the requirement of dual comparison (or mutual unambiguousness as we call it) proposed for node matching algorithms by (Filin and Doytsher 00).

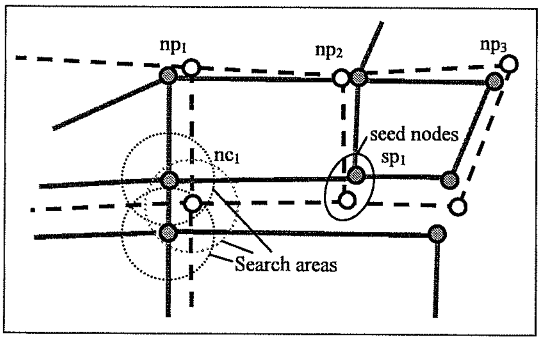


Figure 3. Finding seed nodes (legend see Figure 2)

The principle of the algorithm is illustrated in Figure 3 where one seed node pair (sp_1) can be identified. It has three incident edges showing approximately the same direction. The node pair np_1 does not belong to the seed nodes since the angle difference between one edge pair is too large. The same is true for node pair np_2 because the corresponding nodes do not have the same number of incident edges and for node pair np_3 since the corresponding nodes only have 2 emanating edges. With respect to node cluster nc_1, again no seed nodes can be found: although for both ATKIS nodes only one corresponding node can be detected that lies within the predefined distance and has the same number of incident edges showing the same angles, the condition of bi-directionality or mutual unambiguousness, respectively, is not fulfilled since for the GDF node there is more than one possible matching candidate within the search area.

5.2 Matching nodes and edges

After the seed nodes have been found, different node and edge matching procedures are applied to the data in an iterative fashion. These procedures are based on the principle of calculating similarity measures between potential matching candidates. Therefore, the similarity measures used are first introduced in this section. Then, the algorithm itself is explained in detail.

5.2.1 Calculating similarity measures: In this work, the degree of correspondence or consistency, respectively, between features is determined by evaluating their topological and geometric consistency by appropriate indicators. For nodes, the following similarity values are determined:

- Proximity, determined as the Euclidian distance between two points
- Combined investigation of the number of incident edges (node degree) and the angle differences between emanating edges

The absolute similarity values of both criteria are transferred onto an interval of 0 (no similarity) to 10 (maximum similarity) in order to derive so-called node evaluation values for the proximity indicator (ev_{NP}) and for the combined angle and node degree criterion (ev_{NA}). This mapping of absolute similarity values to evaluation values follows explicit rules which cannot be described here in detail. For example, if the distance between two points is less than 3 m ev_{NP} is 10 and if it is larger than 30 m ev_{NP} is 0. Similarly, ev_{NA} is 10 if the number of incident edges is equal and if the angles of those edges are within a tolerance of 5 degrees, etc. The total node evaluation value (T_{ev} (node)) then is determined as

$$T_{ev}\,(\text{node}) = 0.75 * ev_{NP} + ev_{NA}$$

i.e. the proximity criterion has a little less influence on the total node similarity than the node degree/angle criterion. This kind of weighting results from the experiences made during the investigation of corresponding node candidates. The total node evaluation value is then normalized onto an interval ranging from 0 to 100, with 100 representing the highest similarity. Finally, the normalized value is called the total node similarity value (TS_{NODE}) for a potential node matching pair.

Similarly, the following similarity measures for edges are calculated:

- Length difference, determined as the ratio of length differences to the whole line lengths of both edges
- Angle difference between two edges, determined as the difference between the larger and the smaller angle against the x-axis
- Average line distance, determined as the average distance of the distances of all vertices of two input edges
- Vertex-Hausdorff distance (Davis and Aquino 04), determined as a less complex and easier to compute approximation of the Hausdorff distance (basically calculating the maximum of all minimal distances between two geometries). The Vertex-Hausdorff distance yields either the same results as the regular Hausdorff distance or at least a useful solution in most cases of vector data matching.
- Adjacency relations of start and end nodes, determined as the difference of the number of incident edges of start and end nodes of two edges

Just like for the nodes, each of the absolute edge similarity values is mapped to a corresponding evaluation value by explicit mapping rules, i.e. evaluation values for the length difference (ev_{EL}), the angle difference (ev_{EA}), the line distance (ev_{ED}), the Vertex-Hausdorff distance (ev_{EH}) and the adjacency (ev_{ET}) are derived. For the edges, the scale of the evaluation values ranges also from 0 (no similarity) to 10 (maximum similarity) for each indicator. The different partial evaluation values are finally aggregated into a total edge evaluation value (T_{ev} (edge)) using the following weighted sum approach:

$$T_{ev}\,(\text{edge}) = 3 * ev_{EL} + 3 * ev_{EA} + 2 * ev_{ED} + 4 * ev_{EH} + 4 * ev_{ET}$$

Again, each evaluation value is weighted by a factor which was specified on the basis of the operator's expertise regarding the influence of the different geometric and topological similarity values on the total similarity. The aggregation of similarity measures is a difficult problem that can be further optimized within our approach, for example by using machine learning techniques (see e.g. Bilenko and Mooney 03).

In correspondence to the total node evaluation value, the total edge evaluation value is normalized onto a scale from 0 (no similarity) to 100 (maximum similarity) as well, leading to the total edge similarity value (TS_{EDGE}).

5.2.2 Description of the matching algorithm: After the detection of seed nodes, all edges emanating from corresponding seed nodes are investigated. In the first phase, the end nodes of edges having a similar angle are compared. The basic notion behind this approach is that if two edges have corresponding start nodes and corresponding end nodes, there is a high likelihood that the edges themselves are matching partners. The comparison of the end nodes can lead to the following results:

a. both end nodes are seed nodes, too
   a1. the seed nodes are corresponding
   a2. the seed nodes are not corresponding
b. both end nodes are unmatched
c. one of the end nodes is already matched, the other one is not

After the status of the end nodes has been acquired, the second phase of the algorithm begins. In case (a1), if both end nodes are corresponding seed nodes the probability that the edges are 1:1 counterpart objects as well is very high. In this case, the match between the edges is performed if the total similarity value for the edges exceeds 40. Also in case (a2) a 1:1 match between the edges is possible, but then the total similarity value between them has to be at least 70. Just like the node matches, edge matches are also stored in a list that is continuously growing after each additional iteration.

In case (b), if both end nodes are unmatched, the total similarity value between them is determined. If it is larger than 70 and there is no other matching candidate around that yields a higher similarity value, a match is performed and the nodes are also added to the seed nodes list. The edge matching is then carried out analogously to case (a1). If no match between the end nodes could be established, a 1:1 edge match is only possible if the respective total similarity value is again 70 or above. The same is true for situation (c) in which one end node has already been assigned whereas the other one remained unmatched. This occurs rather seldom. The described procedure is carried out in multiple iterations as long as no new matches can be found.

The 1:1 edge matching is suitable for most cases since the preprocessing step has already split the data sets in a way that during the matching mostly 1:1 matches occur. However, in some cases (see Figure 4) there are still 1:2 matches. Thus, if no 1:1 edge match can be determined, an extended algorithm being able to find 1:2 matches is triggered.

The 1:2 matching algorithm can be illustrated by means of Figure 4. It starts from seed node pair sp_1. The following comparison of the end nodes (gn_1 and an_1) of the incident edges (ge_1 and ae_1) of sp_1 does not lead to a node match and neither does the edge matching procedure, otherwise a 1:1 match could be established. So it is first determined which of the compared edges is the shorter one. The shorter edge ge_1 is then extended in case it has an adjacent edge that – together with ge_1 – shows an angle similar to the angle of the longer edge ae_1. In the example in Figure 4, this is edge ge_2. Eventually, the total similarity between nodes gn_2 and an_1 and edges (ge_1, ge_2) and ae_1 is calculated and if appropriate similarity values can be found, a 1:2 match is established. The principle of the 1:2 match algorithm is adopted from the so-called buffer growing algorithm proposed by (Walter 97).

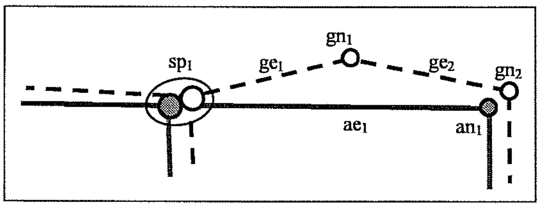


Figure 4. A situation in which a 1:2 edge matching algorithm (legend see Figure 2) is required; the node gn_1 could not be transferred to edge ae_1 during the topological splitting since it lies too far away from it

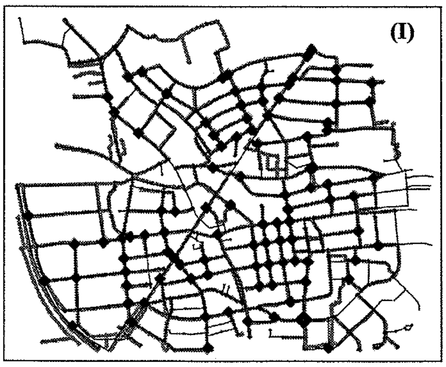
If no new matches can be detected anymore the whole matching follows again applying relaxed constraints. This means that nodes are already accepted as matches if they show a total similarity value of 50 or higher in case no other potential candidate is around, and edge matches are accepted if their similarity is at least 30 in case (a1) or 50 in all other cases.

Also, the seed node detection phase is repeated relaxing the search criteria. Now, all nodes having at least 2 incident edges with corresponding angles (tolerance extended to ± 8 degrees) are added to the seed node list. After the second seed node detection, the whole matching process starts again from the beginning until no new matches can be found anymore.

5.3 Performing Iterations

After the seed node detection, all of the further matching procedures are applied to the data in an iterative way. Thus, starting at the seed nodes the network of matched objects is constantly growing (see Figure 5).

Figure 5 shows different stages of the matching for a test area in the city of Stuttgart, approximately 2 square kilometres in size. In the first stage (I), the seed nodes have been identified (bold dots). The second stage (II) illustrates the situation after 2 iterations of the 1:1 matcher and the third stage (III) displays the final result of the automatic matching procedure. All matched objects are drawn in dark colour/bold style.



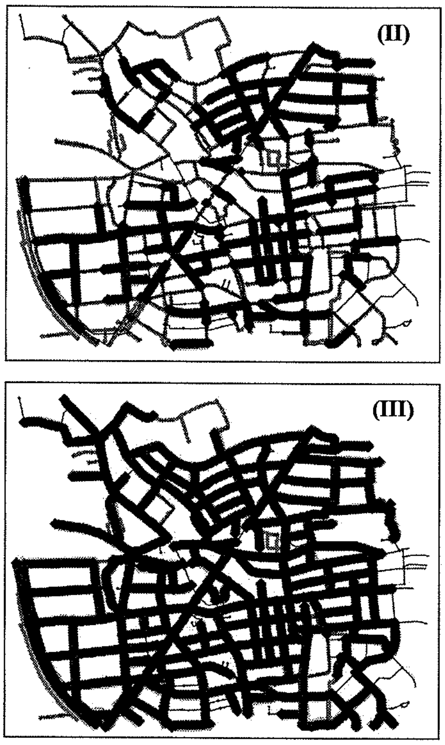


Figure 5: Different stages of the matching process; ATKIS: grey (semi-bold), GDF: black; matched objects are dark grey and bold

5.4 Storing the matches in MRep Relations

According to (Uitermark et al. 96), "geographic data set integration (or map integration) is the process of establishing relationships between corresponding object instances in different, autonomously produced, geographic data sets of a certain region." Basically, these relationships between multiple representations could only be expressed as simple pointers within a bidirectional list, displaying that an object or an object set of data set A can be assigned to an object or an object set of data set B and vice versa ($a_A \leftrightarrow x_B$, $\{c_A, m_A\} \leftrightarrow y_B$, $\{l_A, r_A\} \leftrightarrow \{n_B, s_B\}$, etc.).

However, the relations could also be defined in a more explicit way as it has already been proposed by (Volz and Walter 04). The notion to store explicit relations between multiple representations, so-called MRep Relations, relies on the fact that during matching more information like the mentioned geometric, topological and also attributive similarity measures for corresponding representations can be derived. In our opinion, this additional information can be exploited for multiple purposes (see below).

An example of an MRep Edge Relation that has been established for two multi-representation edges of ATKIS and GDF is illustrated in Figure 6 in the XML-based MultiRepresentational Relation Language (MRRL) exchange format that has been defined within this work. It shows the identifiers of the MRep Relation itself and those of the counterpart objects (source/target_ids), the cardinality of the match, the total similarity measure and the different geometric and topological measures (see 5.2). Similarly, MRep Relations between corresponding nodes can be established. An MRRL file also contains the identifiers of those objects for which no counterpart could be found.

```
<mrepedgerelation>
 <mrepedgerelation_id>mrep_edge_307</mrepedgerelation_id>
  <attributes>
   <general_atts>
    <source_ids>
      <id>atkis_A02MZNE3RE</id>
    </source_ids>
    <target_ids>
      <id>gdf_0x09c658b10c9e11d9901aed0fa2996570</id>
    </target_ids>
    <cardinality>1:1</cardinality>
    <total_similarity>93.75</total_similarity>
   </general_atts>
   <geometric_atts>
    <length_difference>3.00</length_difference>
    <angle_difference>0.05</angle_difference>
    <hausdorff_distance>11.71</hausdorff_distance>
    <avg_line_distance>11.50</avg_line_distance>
   </geometric_atts>
   <topological_atts>
    <startnode_deg_diff>0</startnode_deg_diff>
    <endnode_deg_diff>0</endnode_deg_diff>
   </topological_atts>
  </attributes>
</mrepedgerelation>
```

Figure 6. Excerpt of MRRL, showing an MRep Edge Relation

As it was shown in (Volz 05a), MRep Relations can be used in order to perform a data-driven matching of different geospatial schemas like ATKIS and GDF. By analyzing the MRep Relations and especially the affiliations of corresponding instances to object classes in their source schemas (e.g. to the object class 'Street' in ATKIS or the object class 'Road' in GDF; notice that these affiliations are not displayed in Figure 6), schema similarities or semantic correlations between object classes of disparate schemas, respectively, can be detected. In (Volz 05b) an approach for a shortest path analysis in multi-representation databases was described. Instead of merging multiple representations into one consolidated and consistent data set where only one single representation is available for each real world phenomenon, the proposed technique exploits MRep Relations and thus avoids the conflation process during data analysis.

Furthermore, by introducing similarity measures within MRep Relations, not only the match between corresponding data instances itself can be graphically visualized in a map, but also the degree of correspondence as resulting from the similarity measure can be presented. Thus, an operator can get an overview of the quality of each individual match, i.e. he or she can directly recognize which matches are rather weak and which matches are highly reliable, thereby reducing the manual efforts with regard to the improvement of the automatically produced matching results.

5.5 Dealing with unsolvable cases

If the spatial data to be matched are not very homogeneous, intricate cases can occur that can sometimes even only hardly be solved by human operators, i.e. a complete automatic matching is generally unrealistic. Thus, a software component enabling manual intervention was provided. It allows viewing the correspondences that have been detected by the automatic system. In case the operator disagrees with the result produced automatically, the match can be undone and a new correspondence can be established. Of course, all the cases which the operator detects as matches and which have not been recognized during the automatic process can be solved as well.

6. RESULTS OF THE AUTOMATIC MATCHING

Generally, the approach shows a good performance. In the test area depicted in Figure 5, 98.86% of all the 985 ATKIS objects and 97.31% of the 1003 GDF objects were matched correctly or were correctly recognized as single representations having no counterpart in the other data set. The percentage values result from a comparison of the automatic approach and a reference matching carried out by a human operator. The whole automatic matching process for the test area takes less than 9 seconds. However, some problems were encountered that were on the one hand related to the topological splitting during the pre-processing step, on the other hand they were caused by the algorithm itself (see Figure 7). These problems shall be discussed here.

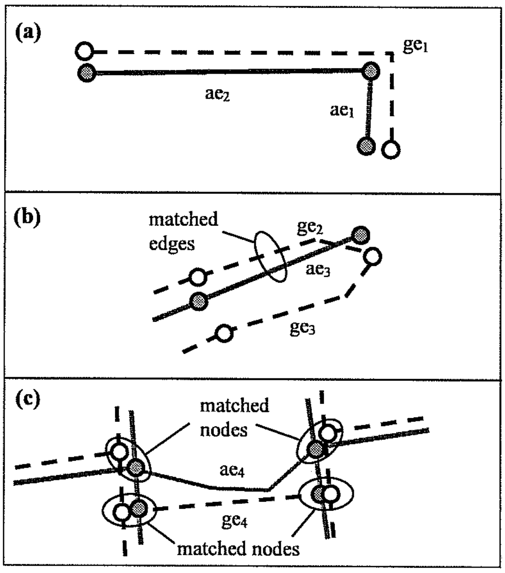


Figure 7. Some problems of the matching algorithm (legend see Figure 2)

During the topological splitting in the pre-processing phase, only those edges which are roughly parallel are considered for node transfers. However, this leads to problems in some cases, as it is depicted in Figure 7 (a) where edge ge$_1$ (total angle approximately 160 degrees) exceeds the angle tolerance both to edges ae$_1$ (total angle approximately 90 degrees) and ae$_2$ (total angle approximately 180 degrees) although the *segments* of ge$_1$ are almost parallel to ae$_1$ and ae$_2$. Thus, no new node is introduced in ge$_1$ and consequently no matches can be detected. On the other hand, if the criterion of parallelism is dropped, in some situations nodes are generated where this is actually not correct. Thus, more indicators have to be introduced during the topological splitting allowing for a better performance of the algorithm.

In another situation shown in Figure 7 (b), one GDF road is represented as two topologically separate edges (ge$_2$ and ge$_3$), whereas in the ATKIS data set it only consists of one line (ae$_3$). During the automatic matching proposed here, only correspondences between ae$_3$ and ge$_2$ can be found whereas edge ge$_3$ remains unmatched. However, if the matching was correct, ae$_3$ would be related both to ge$_2$ *and* ge$_3$. This type of matches can up to now not be recognized automatically by our algorithm which brings up another task for future improvements. One approach to handle this kind of problem could be to consider also thematic attributes like street names during the matching.

Another drawback of the approach is illustrated in Figure 7 (c). Here, all edges displayed are matched except ae$_4$ and ge$_4$, although they are very likely corresponding. The reason for the error is that both the start and the end nodes of the edges are not corresponding, and consequently the approach fails (notice that corresponding edges have to have either corresponding start or corresponding end nodes, otherwise no edge matches are triggered). In order to solve such problems in further versions of the approach, edges which remained unmatched could be investigated after the iterative matching has been carried out and similarity measures could be calculated for potential candidates. Such mechanisms have not been incorporated in the implementation yet.

Moreover, in very few cases, further errors and inadequate matches, respectively, have been introduced in the test area due to the relaxation of matching constraints, i.e. there might be a potential for optimization, too.

It has to be mentioned that the test area did not contain many difficult cases. Therefore, we are planning to apply the algorithm to test scenes which are more complicated so that we can discover further error sources and receive hints on potential improvements of our approach in order to make it as generic as possible.

7. CONCLUSION AND OUTLOOK

In this paper we have proposed an automatic matching algorithm for linear street data of ATKIS and GDF. It relies on a pre-processing of the data and combines iterative node and edge matching concepts based on the detection of similarity measures. The results of the process are stored explicitly as relations between multiple representations. The matching basically provides reliable results. In some cases (see previous section), however, the approach still fails to recognize the situation correctly and establishes wrong matches or misses available correspondences. Thus, an improvement of the pre-processing step and of the matching algorithm itself with respect to the drawbacks found is one of our future goals. Also, more complex situations have to be considered that will probably reveal more problems and can lead to a further enhancement of the procedure. Additionally, we plan to apply the approach to the matching of multi-scale linear street data (1:250000 and 1:25000) which probably requires different methods for generating similarity values. In principle, the presented matching algorithm could be used for identifying corresponding linear features of other object types (e.g. hydrological networks) which has to be investigated as well. Since the matching is up to now merely geometry- and topology-based, including semantic aspects will definitely be a point to look at with respect to future improvements. Another issue will deal with the merging or conflation of multiple representations based on MRep Relations, also considering data quality/uncertainty issues.

8. REFERENCES

ADV: Homepage of the Working Committee of the Surveying Authorities of the States of the Federal Republic of Germany, http://www.adv-online.de, accessed: 11.2004

Aho, A., Hopcroft, J. E., Ullman, J. D.: Data Structures and Algorithms. Addison-Wesley Series in Computer Science and Information Processing, (1987), 427 p.

ANZLIC: Homepage of the Spatial Information Council for Australia and New Zealand, http://www.anzlic.org.au, accessed: 04.2005

Beeri, C., Doytsher, Y., Kanza, Y., Safra, E., Sagiv, Y.: Finding Corresponding Objects when Integrating Several Geo-Spatial Datasets. In: Proceedings of the 13th ACM International Workshop on Geographic Information Systems, Bremen, Germany, (2005), pp. 87-96.

Bilenko, M., Mooney, R.J.: Employing Trainable String Similarity Metrics for Information Integration. In: Proceedings of the IJCAI-2003 Workshop on Information Integration on the Web, Mexico, (2003), pp. 67-72.

Bishr, Y. A., Pundt, H. Rüther, C.: Proceeding on the Road of Semantic Interoperability - Design of a Semantic Mapper Based on a Case Study from Transportation. In: Včkovski, A., Brassel, K.E., Schek, H.-J. (eds.): Proceedings of the 2nd International Conference on Interoperating Geographic Information Systems, Zurich, Lecture Notes in Computer Science, Heidelberg, Berlin, (1999), pp. 203-215.

Bofinger, J.M.: Analyse und Implementierung eines Verfahrens zur Referenzierung geographischer Objekte. Diploma Thesis at the Institute for Photogrammetry, University of Stuttgart, (2001), 76 pages.

Cobb, M., Chung, M., Miller, V., Foley, H., Petry, F., Shaw, K.: A Rule-Based Approach for the Conflation of Attributed Vector Data. GeoInformatica 2(1), (1998), pp. 7-35.

Davis, M., Aquino, J. : Java Conflation Suite (JCS), Technical Report, (2003), 48 p. Acces via: http://www.jump-project.org/, accessed: 02.2002

Dunkars, M. : Matching of Datasets. In: Proceedings of the 9th Scandinavian Research Conference on Geographical Information Science (SCANGIS) '03, June 4th to 6th Espoo, Finland, (2003), pp. 67-78.

FGDC: Homepage of the Federal Geographic Data Committee of the United States, http://www.fgdc.gov, accessed: 04.2005

Filin, S., Doytsher, Y.: Detection of Corresponding Objects in Linear-Based Map Conflation, Surveying and Land Information Systems, Vol. 60, No. 2, (2000), pp. 117-128.

Gösseln, G. v., Sester, M.: Integration of Geoscientific Data Sets and the German Digital Map using a Matching Approach. In: Proceedings of the XXth ISPRS Congress, Comm. IV, Istanbul, Turkey, (2004), pp. 1249-1254.

GSDI: Homepage of the Global Spatial Data Infrastructure Association, http://www.gsdi.org/, accessed: 04.2005

Jones, C. B., Kidner, D. B., Luo, L. Q., Bundy, G. L., J. M. Ware: Database Design for Multi-Scale Spatial Information System. Int. J. Geographical Information Science 10(8), (1996), pp. 901-920.

JUMP: Java Unified Mapping Platform, http://www.jump-project.org/, accessed: 11.2005

Kraft, W.: Entwurf von Zuordnungsalgorithmen zur Fortführung und Überprüfung von raumbezogenen Datenbeständen. Diploma Thesis at the Institute for Photogrammetry, University of Stuttgart, (1995), 75 pages.

Kraut, M.: Zuordnung und Conflation heterogener Straßendaten. Diploma Thesis at the Institute for Photogrammetry, University of Stuttgart, (2003), 109 pages.

Mantel, D., Lipeck, U.: Matching Cartographic Objects in Spatial Databases. In: Proceedings of the XXth ISPRS Congress, Comm. IV, Istanbul, Turkey, (2004), pp. 172-176.

Nexus: Homepage of the Nexus Project of the University of Stuttgart, http://www.nexus.uni-stuttgart.de, accessed: 11.2005

Pandazis, J.: TR 4011 EVIDENCE - Final Report, (1999), Brussels.

Stigmar, H.: Matching Route Data and Topographic Data in a Real-Time Environment. In: Proceedings of the 10th Scandinavian Research Conference on Geographical Information Science (SCANGIS) '05, June 13th to 15th Stockholm, Sweden, (2005), pp. 89-107.

Uitermark, H.: The Integration of Geographic Databases. Realising Geodata Interoperability through the Hypermap Metaphor and a Mediator Architecture. In: Rumor, M., McMillan, R., Ottens, H.F. (eds.): Proceedings of the 2nd Joint European Conference & Exhibition on Geographical Information (JEC-GI) '96, Vol. I, Barcelona, (1996), pp. 92-95.

Van Wijngarden, F., van Putten, J., van Oosterom, P., Uitermark, H.: Map Integration – Update Propagation in a Multi-Source Environment. In: Proceedings of the 5th ACM international workshop on advances in geographic information systems, Las Vegas, Nevada, United States, (1997), pp. 71-76.

Volz, S.: Data-driven Matching of Geospatial Schemas. In: Cohn, A.G., Mark, D.M. (eds.): Spatial Information Theory. Proceedings of the International Conference on Spatial Information Theory (COSIT '05), Ellicottville, NY. Lecture Notes in Computer Science 3693, (2005a), pp. 115-132 .

Volz, S.: Shortest Path Search in Multi-Representation Street Databases, In: Proceedings of the 3rd Symposium on Location Based Services and TeleCartography, Vienna, Austria, (2005b), pp. 125-130.

Volz, S., Walter, V.: Linking Different Geospatial Databases by Explicit Relations. In: Proceedings of the XXth ISPRS Congress, Comm. IV, Istanbul, Turkey, (2004), pp. 152-157.

Walter, V.: Zuordnung von raumbezogenen Daten – am Beispiel der Datenmodelle ATKIS und GDF. Dissertation, Deutsche Geodätische Kommission (DGK), Reihe C, Heft Nr. 480, (1997), 127 pages.

Walter, V., Fritsch, D.: Matching Spatial Data Sets: a Statistical Approach, Int. J. Geographical Information Science 13(5), (1999), pp. 445-473.

Weis, M., Naumann, F.: Detecting Duplicate Objects in XML Documents. In: Proceedings of the SIGMOD International Workshop on Information Quality in Information Systems (IQIS) '04, Paris, (2004), pp. 10-19.

Xiong, D., Sperling, J.: Semiautomated Matching for Network Database Integration, ISPRS Journal of Photogrammetry and Remote Sensing, Special Issue on Advanced Techniques for Analysis of Geo-spatial Data, Volume 59 (1-2), (2004), pp. 35-46.

Zhang, M., Shi, W., Meng, L.: A Generic Matching Algorithm for Line Networks of Different Resolutions. In: Proceedings of the ICA workshop on generalisation and multiple representation, A Coruña, Spain, (2005).

Acknowledgements

The research presented here is part of the Nexus project which is supported as a Center of Excellence called "SPATIAL WORLD MODELS FOR MOBILE CONTEXT-AWARE APPLICATIONS" under grant SFB 627 by the Deutsche Forschungsgemeinschaft (DFG - German Research Council).

The test data have kindly been provided by the Navteq company (GDF) and the state survey office of the federal state of Baden-Wuerttemberg (ATKIS).

The author likes to thank the anonymous reviewers for their valuable comments.