Link based Conflation of Geographic Datasets

Jan-Henrik Haunert
Institute of Cartography and Geoinformatics
University of Hannover
Appelstraße 9a, 30167 Hannover, Germany
jan.haunert@ikg.uni-hannover.de

KEY WORDS: MAP CONFLATION, MULTIPLE REPRESENTATION, UPDATE, GIS

1. Introduction

The conflation of geographic data from different sources is an important research area in geographic information sciences. In recent time this is increasingly motivated by the demand for a higher usability of expensively collected data on the part of the national mapping agencies. For this reason also the linkage of geographic datasets with different scales and their mutual administration in one database is under examination. Linking objects of different scales will ease the propagation of new information to a generalized dataset and the transfer of information from lower to higher scale becomes more feasible.

In this paper a method for a topological consistent transfer of features is presented, which assumes, that links between corresponding features exist. The method is based on a map transformation which considers locally existing geometric differences between the two involved datasets. Even though algorithms and mathematical formulas of similar existing methods differ among each other and in relation to the presented work, the term “rubber sheeting” is often used for this kind of transformation and will also be used here. It refers to the illustration of the map as a flexible membrane that is fitted into a frame and hence is forced to become deformed. Rubber sheeting can be implemented by an interpolation of coordinate differences that are given for control points. In this work an infinite number of control points on corresponding linear elements is defined.

After an overview on related work, this paper addresses how the transformation can be set up by exploiting links between linear features. Also the mathematical foundations for the transformation are explained and results for existing datasets are shown. In the conclusion limits are presented that require further development and research.

2. Related work

2.1. Context of this work at the ikg

In the project “WIPKA” the Information Systems Institute of the University of Hannover, the Federal Agency for Cartography and Geodesy and the ikg have built up a database, which is capable of keeping multiple representations of geographic objects (Multiple Representation Database, MRDB). It has been designed for the four digital topographic datasets ATKIS, which are orientated on the content of the topographic maps at the scales 1:25.000 (Basis DLM), 1:50.000 (DLM 50), 1:250.000 (DLM 250) and 1:1.000.000 (DLM 1000). The datasets were linked, based on features that correspond to each other (Mantel & Lipeck, 2004).

Data of the same geographic objects was collected independently for the different representations in the past. Datasets differ in completeness and up-to-dateness. The aim is now, that information that is available in at least one representation, but missing in the others,

becomes available in all layers of the MRDB. Another demand is to propagate new information which is inserted on any layer to the other datasets (Anders, 2004).
In another project, which is funded by the German Research Foundation (DFG) solutions for these tasks are searched. The presented work is part of this project and shows how geometric differences between linked datasets can be handled during the propagation of features. The method is not capable of overcoming big scale-differences between two datasets. Therefore, for the propagation of updates in an MRDB it can only be used within an approach of two steps. Firstly, the scale of the features needs to be adapted to the target dataset by generalization methods. Secondly, the presented rubber sheeting algorithm can be performed to ensure a topological correct integration of the updates to the target dataset.

2.2. Related work concerning map conflation

Generally, the conflation of different datasets is separated into two tasks: Identifying correspondences and aligning matches via rubber-sheeting. Often, these steps are processed iteratively as in (Lupien & Moreland, 1987) and (Saalfeld, 1988).
In this work, the first task is assumed to be accomplished to a certain stage. Correspondences between features are given as links, which support the search for point matches. However, it has to be pointed out, that matching is often very complex and research on this issue is still going on. Successes were made by Rosen & Saalfeld (1985), Walter & Fritsch (1999) and Uitermark (2001).
Usually, for the map transformation an interpolation of coordinate differences is defined, which allows the calculation of corrections to the coordinates of all points within a defined part of the map. By applying these corrections the map transformation is performed.
Normally, the interpolation function is defined on a triangulated network of the matched points (Gillman, 1985) or by a weighted mean over a set of matched points. Herein the weights are expressed by a function of the distance between the query point and the matched points (Doytsher et al., 2001).
Gielsdorf & Gründig (1997) use an algorithm, based on a least square adjustment on a TIN, which physically models the characteristics of a flexible membrane. For this algorithm the term rubber-sheeting fits exactly, because the map plane behaves exactly like a sheet of rubber under deformation.

3. Problem definition

Two datasets that represent overlapping areas of the real world are given. Both contain networks of linear elements such as roads, rivers, railways and city boundaries. Also links between pairs of features are given in a table. These links state that both features represent at least partly the same real world object.
The delineation of the features does not necessarily have to be identical in both datasets. For example, in one dataset the intersection of a road and a river splits both into two features, while the intersection is not accounted in the other dataset. Also a road section between two junctions might be represented by two features in one dataset, while it is represented by three features in the other. Generally, the cardinality of the links can be m:n.
Furthermore, both datasets may contain singletons, what means that a feature is not linked or in other words does not have a corresponding feature in the other dataset. When dealing with generalized data, the consideration of this aspect is inevitable, because elements with a lower importance might have been eliminated (Figure 1, left box).
The first step of the algorithm is based on points whose correspondences are well defined by the links and the topology, such as junctions in the road network. If these points do not exist, the algorithm will fail. However, for the road datasets of an urban area this is not a problem.

Topologic discrepancies are not considered. A four-intersection cannot be represented as two three-intersections in the other dataset (Figure 1, right box). This restriction generally applies to rubber-sheeting: A split of the membrane is not allowed. Consequences of this limitation will be discussed in section 6.
The problem which is approached in the next section is the topologically consistent transfer of a feature from one dataset into the other. This applies to features that are attached to elements of the network, such as house-numbers, which are fixed to a road, as well as features that have their position within a mesh of the network.

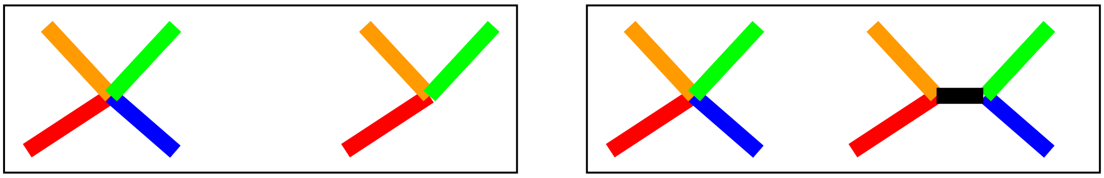

Figure 1) Topologic differences of networks that are allowed (left box) and not allowed (right box) for the developed conflation method (same colors represent corresponding features).

4. Methodology

4.1. Overview

The initial stage of linked features is depicted in Figure 2a. The algorithm is divided into three steps:

1. Links and the topology of the network are analysed to gain coordinate differences in points for which correspondences are well-defined (Figure 2b).
2. Additional points are inserted in both datasets to branches between pairs of the previously matched points, so that every point of a linked feature can be matched to a corresponding point (Figure 2c).
3. For a point that should be transformed the containing face of linked features is searched in the source dataset. Corrections are interpolated from coordinate differences that are given for the endpoints of the line elements bounding this face. The correction of a point that falls on a matched line segment can be calculated by a linear interpolation of differences in the endpoints of the line.

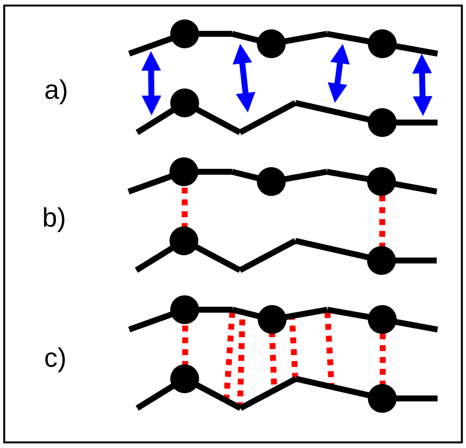

Figure 2) Stages of the algorithm. Arrows show feature based links. Dashed lines show correspondences of points. The black circles show the endpoints of linear features.

4.2. Finding well-defined point correspondences from topology and links

For many points the correspondences are implicitly given by the links and the topology of the network. A candidate for this category is every point in the source datasets (dataset 1) in which more than one linked feature meet. The corresponding point in the other dataset (dataset 2) can be found by geometric union and intersect operations. For every feature in dataset 1, which is not a singleton and which intersects the point in dataset 1, the geometric union of all linked geometries in dataset 2 is applied. The intersection of all these unions returns a distinct set of geometries which represent possible correspondences in dataset 2.

Normally the set contains one single point. In this case the correspondence can be assigned immediately. Other cases that occur are a one-dimensional geometry and a higher number of geometries. The first means, that the corresponding point in dataset 2 is somewhere within the linear result. The correspondence is not well-defined from the topology and the links and will be defined in step 2. The second case happens, when the linked features intersect in more than one point. These cases are very rare and can be postponed to the end of the first step. The correct correspondence can be found by a topology check of the possible combinations from these ambiguous results.
Theoretically, these rules are not complete. Special cases in which only parts of a branch between two junctions are omitted in one network are not considered, but could be handled with additional rules.
Dead-ends are not assigned as corresponding points, even though the correspondence of roads that lead to these points might be of the cardinality 1:1. This restriction is made, because it is likely that a road is cut off earlier in one dataset than in the other.

4.3. Defining additional control points

For all points of both dataset, which were not assigned to correspondences in the first step, new points are inserted in the other dataset. For a point which lies on a branch between two control points a point is simply added at the same ratio of the total length in the corresponding branch. For this work it was sufficient, because the density of junctions or control points is high in the test dataset, which represents an urban area. Problems might occur in rural areas, where junctions are rare, but roads have a lot of curves. Here it should be considered to apply a geometric comparison of road sections to find better correspondences. Also the approach will not be sufficient if lines are simplified to a high degree in one dataset due to scale differences.
For additional points on dead ends, which are only at one end incident to a control point, the preservation of the absolute distance from this point could be postulated as the simplest rule.

4.4. Transformation of the interior of a face

After step 2 for every point P_i, $i = 1 \ldots n$ of a face, that is delimited by linked features, differences Δx_i, Δy_i are given. These values can be used for the interpolation of geometric corrections which have to be added to the coordinates of an interior point P_0. The interpolation function on which this work builds up results from the sum over the differences on the boundary which are weighted by the reciprocal of the squared distances between a boundary point and the interior point P_0. A detailed documentation of its application can be found in Doytsher et al. (2001). Formulas will be shown for x-components only. The calculation of the y-components is carried out analogously.

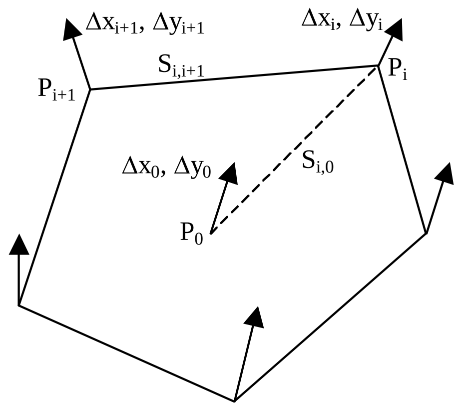

Figure 3) Interpolation of coordinate differences

$$\Delta x_0 = \sum_{i=1}^{n} \left(\frac{1}{S_{i,0}^2} \cdot \Delta x_i \right) \Bigg/ \sum_{i=1}^{n} \frac{1}{S_{i,0}^2} \quad (4.1)$$

To overcome an inhomogeneous distribution of points on the boundary, these points are often densified by linear interpolation. With formulas (4.2) and (4.3) the coordinates and corrections of a point at the distance s from the origin of the line segment can be calculated.

$$x(s) = x_i + s \cdot (x_{i+1} - x_i)/S_{i,i+1} \qquad (4.2) \qquad\qquad \Delta x(s) = \Delta x_i + s \cdot (\Delta x_{i+1} - \Delta x_i)/S_{i,i+1} \qquad (4.3)$$

The distance from the query point P_0 to the interpolated boundary point results from

$$S_0^2(s) = (x_i - x_0 + s \cdot (x_{i+1} - x_i)/S_{i,i+1})^2 + (y_i - y_0 + s \cdot (y_{i+1} - y_i)/S_{i,i+1})^2 . \qquad (4.4)$$

The approach in this work goes one step ahead by defining an infinite density of interpolated points on the boundary. By this it is possible to replace the sum over the contributions from a finite number of points by the integral over the boundary. This results in a sum over all edges whose contributions are calculated by the definite solution of the integral.

$$\Delta x_0 = \sum\left(\int_0^{S_{i,i+1}} \frac{\Delta x(s)\cdot ds}{S_0^2(s)}\right) \Bigg/ \sum\left(\int_0^{S_{i,i+1}} \frac{ds}{S_0^2(s)}\right) = \sum\left(S_{i,i+1} \cdot \int_0^1 \frac{\Delta x(r)\cdot dr}{S_0^2(r)}\right) \Bigg/ \sum\left(S_{i,i+1} \cdot \int_0^1 \frac{dr}{S_0^2(r)}\right) \qquad (4.5)$$

The solutions of these integrals can be found in Bronstein & Semendjajew (1967, p. 37, integrals 40 and 44). The advantage of this method compared to the original algorithm is, that the result is independent from the distribution of points on the boundary. The number of summands remains small, since no additional boundary points have to be interpolated. Figure 4 (left) shows that the simple weighted sum does not suffice when the distribution of given coordinate differences is inhomogeneous. A better result can be created with the developed method (Figure 4, right).

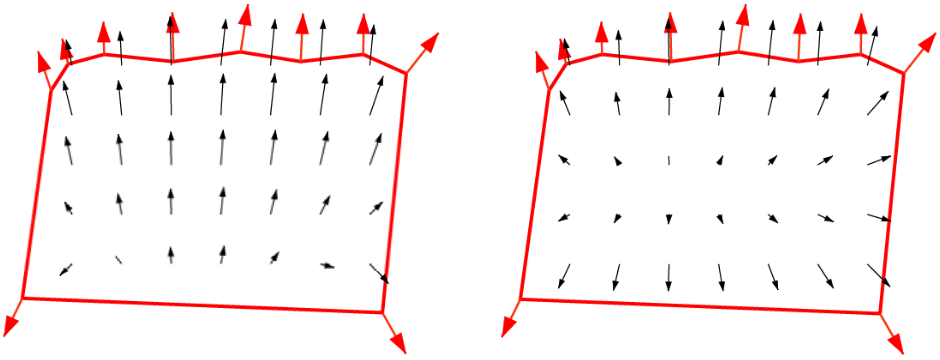

Figure 4) Results for an inhomogeneous distribution of control points with given coordinate differences (arrows on boundary) – achieved with formula 4.1 (left) and with formula 4.5 (right).

5. Results

For a test of the algorithm data from the ATKIS DLM 250 and ATKIS DLM 1000 was used (Figure 5). Both datasets contain roads, rivers and railways that were linked according to the requirements in section 3.
Since data from the DLM 1000 is used by military users for aerial navigations, the positions of towers have been collected for this dataset.
Even though the specifications of the DLM 250 define that these objects have to be represented, the information is missing in this dataset, due to different updating cycles. This is an interesting case when the low resolution dataset contains more detailed information than the next detailed one. Thus, it is aimed to support the DLM 250 with the available information from the dataset with the lower resolution.
Figure 5a shows both datasets for the city center of Hannover before the propagation has happened. The networks correspond to each other without topological conflicts.

In this example it would be possible to add both towers at their original position to the other dataset without violating their topological relationships. Generally, this does not apply. An example of a point for which the containing mesh would change if no transformation is applied is marked with an x in Figure 5a.
However, also for the two towers the transformation should be applied. The tower which is close to the river is a part of the mediaeval fortification and forms a famous historical landmark. Its relative closeness to the river should be preserved in any case.
Figure 5b shows the difference vectors which have been constructed in every vertex of the networks. Finally, figure 5c shows the vector field of point corrections which is defined by the differences on the bounding features of the face and the interpolation function. The towers were transformed and added to the map at the proper location.

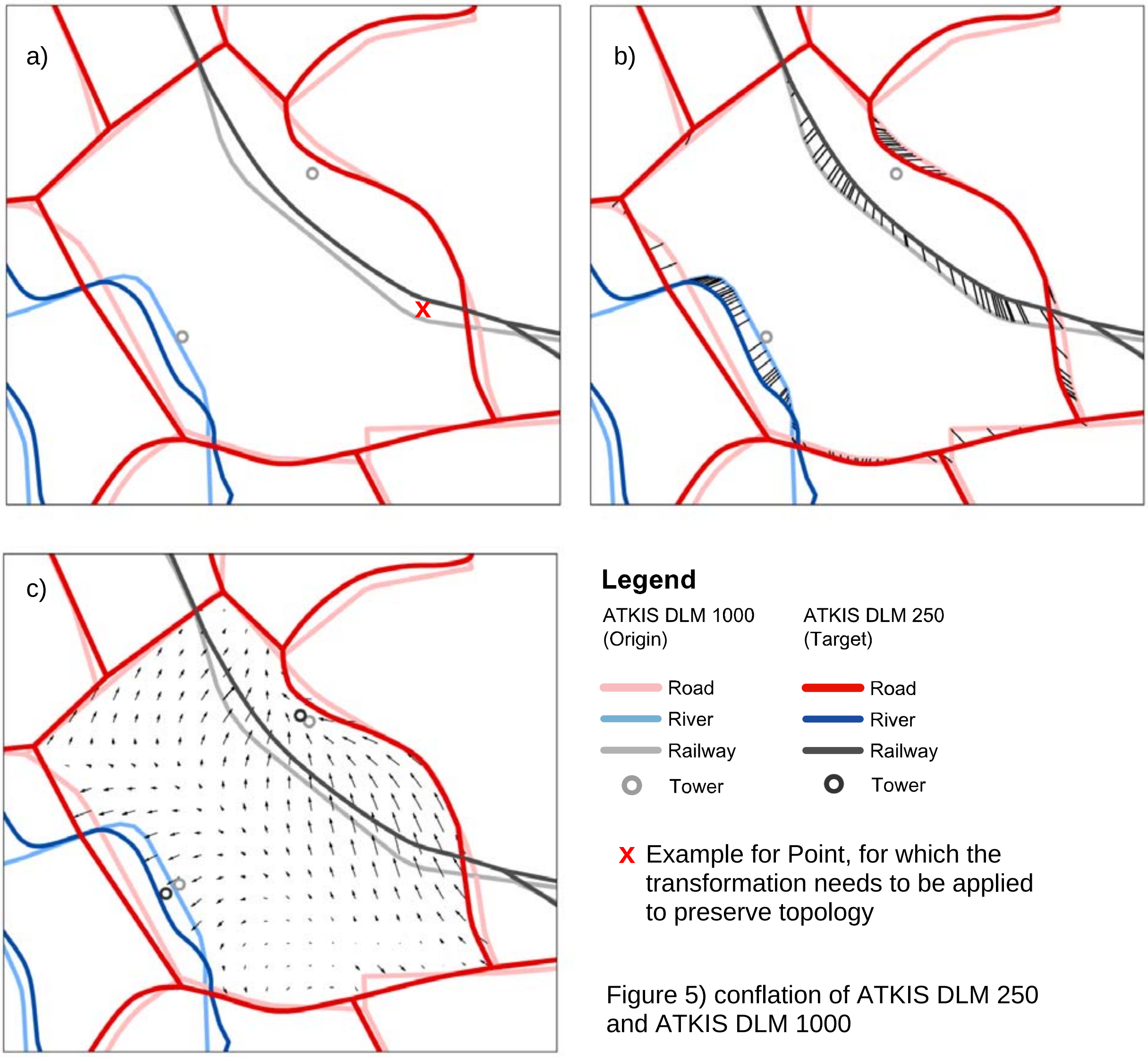

Figure 5) conflation of ATKIS DLM 250 and ATKIS DLM 1000

6. Conclusion and Future Work

In conclusion it can be stated, that existing links between corresponding features are a great support for the classic map conflation problem. The task of matching points can be simplified to a high degree.
For the rubber sheeting an interpolation function was applied which rather considers contributions of lines than contributions of points. By this the dependency of the result from the distribution of control points on the boundary of a face was overcome. However, if

additional constraints among the objects have to be taken into account (e.g. rectangular angles) methods like the one proposed by Gielsdorf & Gründig have to be used.
The exclusion of topologic discrepancies between the matched networks is a fundamental restriction of the method. In generalised geographic networks these discrepancies often occur due to typification. Overcoming this limitation is a very difficult task. For the propagation of information from higher to lower scale an approach of two steps could be developed if the generalisation rules were known. The first step would contain a generalisation of the road network in which the added information is considered. In this step a topology needs to be created which is equivalent to the topology of the target dataset. In the second step a rubber-sheeting transformation could be applied to overcome remaining geometric differences of the datasets. Herein the results of this work could be used.
For the propagation of information to a dataset with a higher scale, which was suggested in this paper, appropriate representations of the transferred features need to be developed. Generally, it will not be possible to reverse the generalization process. A specialization can not be reached. Therefore, the features which were transferred by this method might be of lower quality than those which were originally collected for the high resolution dataset. For the user, the information about the origin of this data should be provided with additional meta data. The development of an appropriate meta data schema for this purpose needs further development.

7. Acknowledgement

This work shows results from the project entitled Updating of Geographic Data in a Multiple Representation Database. The project is funded by the German Research Foundation (Deutsche Forschungsgemeinschaft, DFG). It is part of the bundle-project entitled Abstraction of Geographic Information within the Multi-Scale Acquisition, Administration, Analysis and Visualisation.

8. References

Anders, K.-H., 2004: MRDB Approach for Automatic Incremental Update – ICA Workshop on Generalisation and Multiple Representation; 20-21 August 2004 – Leicester. http://ica.ign.fr/Leicester/paper/Anders-v1-ICAWorkshop.pdf, accessed February 2005.

Bronstein, I. N. & Semendjajew, K. A., 1967: Taschenbuch der Mathematik. Verlag Harri Deutsch, Zürich.

Doytsher, Y., Filin, S., Ezra, E., 2001: Transformation of Datasets in a Linear-based Map Conflation Framework. Surveying and Land Information Systems 61(3), pp. 165-175.

Gielsdorf, F., Gründig, L., 1997: Nachbarschaftstreue Anpassung auf der Basis des Membranmodells – Zeitschrift für Vermessungswesen 5/1997.

Gillman, D. W., 1985: Triangulations for Rubber-Sheeting. Proceedings of AUTOCARTO 7; 11-14 March 1985 – Washington D.C., pp. 191-199.

Lupien, A. E. & Moreland, W. H., 1987: A General Approach to Map Conflation. Proceedings of AUTOCARTO 8; 29-3 April 2 – Baltimore, Maryland, pp. 630-639.

Mantel, D. & Lipeck, U. W., 2004: Matching Cartographic Objects in Spatial Databases, in: Int.Archives of Photogrammetry, Remote Sensing and Spatial Inf.Sciences, Vol. 35, Part B4, pp. 172-176, ISPRS, Istanbul, 2004.

Rosen, B. & Saalfeld, A., 1985: Match Criteria for Automatic Alignment. Proceedings of AUTOCARTO 7; 11-14 March 1985 – Washington D.C., pp. 456-462.

Saalfeld, A., 1988: Automated map compilation. International Journal of Geographical Information Science 2(3), pp. 217-228.

Uitermark, H., 2001. Ontology-based geographic data set integration. Doctor thesis, Deventer, Netherlands.

Walter, V. & Fritsch, D., 1999: Matching spatial data sets: a statistical approach. – International Journal of Geographical Information Science 13(5), pp. 445-473.