HIERARCHICAL QUALITY INSPECTION OF SPATIAL DATA BY DATA INTEGRATION

Hainan Chen
Volker Walter
Institute for Photogrammetry
Universitaet Stuttgart
Geschwister-Scholl-Str. 24D
D-70174 Stuttgart
hainan.chen@ifp.uni-stuttgart.de
volker.walter@ifp.uni-stuttgart.de

ABSTRACT

In this paper we introduce an approach for quality inspection based on map matching and data integration. Two datasets are the input for the process. At first, the edges of the datasets are matched manually. Then, the form of each matching pair is determined and corresponding nodes are matched. In the final step, quality measurements on different levels of granularity are calculated. On the dataset level, a global geometric quality measure based on the evaluation of adjacency matrices is proposed. Furthermore, completeness and topologic similarity are measured. On the matching pair level, geometric modeling, geometric similarity and attribute similarity are analyzed.

INTRODUCTION

Spatial data are collected by different institutions for different purposes which lead to multiple representations of the same objects of the world. Multiple representations mean that redundant information is available which can be used for the evaluation and improvement of the quality of the data. In the following we describe an approach for quality inspection based on map matching and data integration. The approach can be applied for large datasets with similar scales (e.g. GDF/NavTeq, GDF/TeleAtlas, OpenStreetMap and ATKIS) and considers not only the geometry of the data but also the attributes and the topological relations. For datasets with strong different scales, a generalization process can be applied as a preprocessing in order to achieve a similar scale.

The remaining paper is structured as follows. After a discussion of existing work, the used matching model is introduced. Then, the automatic recognition of the form of matching pairs and the automatic node matching are explained in detail. The approach of hierarchical quality inspection is presented subsequently. A discussion of the approach and a summary conclude the paper.

RELATED WORK

Spatial data quality has been a topic of intensive research for several decades. Different quality models and quality characteristics have been developed. For example, ISO 19113 recommends the use of the following quality characteristics for quality inspection (ISO19113, 2002): *Logical Consistency*, *Completeness*, *Geometric Accuracy*, *Thematic Accuracy* and *Temporal Accuracy*.

ISO 19114 divides quality inspection into three classes: direct internal, direct external and indirect methods (ISO19114, 2003). Direct internal methods check the data by inspecting the Logical Consistency of a dataset. Logical Consistency represents the degree of adherence to the logical rules of data structure, attribution and relationships and can be measured automatically without other external information (Caspary & Joos, 1998).

Direct external methods evaluate a dataset by comparing the data elements with a reference data set. High resolution satellite and aerial pictures are typically used for this kind of quality inspection (Becker et al., 2008; Gerke et al., 2004; May, 2002). The quality characteristics *Completeness*, *Geometric Accuracy, Thematic Accuracy* and *Temporal Accuracy* can be checked for data elements which can be extracted from the images. User Generated Content (UGC) or Volunteered Geographic Information (VGI) can also be used for direct external inspection (Franz, 2008; Otto, 2005; Visintainer et al., 2008). *"This network of human sensors has over 6 billion components, each an intelligent synthesizer and interpreter of local information"* (Goodchild, 2007).

Indirect methods do not evaluate the data elements itself but evaluate meta-information about a dataset. For example, ontologies can be evaluated for deriving quality information (Frank, 2008).

In this paper we propose the integration of different vector datasets for quality evaluation of spatial data which is a direct external quality inspection. The first step of data integration is the identifying of correspondences in the datasets (matching). A "Buffer Growing" method for solving this matching problem was first described in (Walter & Frisch, 1999). The matchings are subdivided into 1:1, 1:n and n:m matchings. This approach was extended by Zhang and Meng (2006) with unsymmetrical buffers. An iterative approach for "Buffer Growing" for data integration on schema level is presented in (Volz, 2006). An approach for matching of linear objects in raster data was developed by Seo and O'hara (2009).

Different map conflation approaches have been developed to improve the data quality of spatial data sets. A hierarchical rule-based system for conflation considering the data quality and the map scales of the data sources was introduced in (Cobb et al., 1998). A map conflation approach for datasets with different resolutions was developed in (Edwards & Simpson, 2002). An automatic map conflation approach for unmatched objects is described in (Deretsky & Rdony, 1993). Mathematical methods (for example Rubber Sheeting or special geometric transformations) for map conflation are described in (Casado, 2006; Haunert, 2005). The solving of conflicts due to different geometric modeling in spatial datasets is discussed in (Haunert, 2005).

MATCHING

In our study we use two different datasets (GDF/TeleAtlas and GDF/NavTeq) which were collected by different companies and at different points in time. Since the two datasets are developed for the same application (Navigation) and use the same data model (GDF – Geographic Data File (ISO 14825, 2004)), many redundancies are available, which can be used for quality inspection. Two test areas are selected for this study. Test area I is located in downtown city area in Stuttgart and test area II in a rural area in Öhringen – both located in the Southern part of Germany. **Table 1** summarizes the number of edges and nodes in the two test areas.

Table 1: Description of test areas

Dataset	Test area I (Urban Area 2×2 km)		Test area II (Rural Area 5×6 km)	
	Edges	Nodes	Edges	Nodes
NavTeq (Q1/2005)	1293	948	690	582
TeleAtlas (Q1/2006)	1713	1246	1991	1537

Figure 1 gives an overview of the two areas (NavTeq in solid and TeleAtlas in dashed). Differences between the two datasets can be observed.

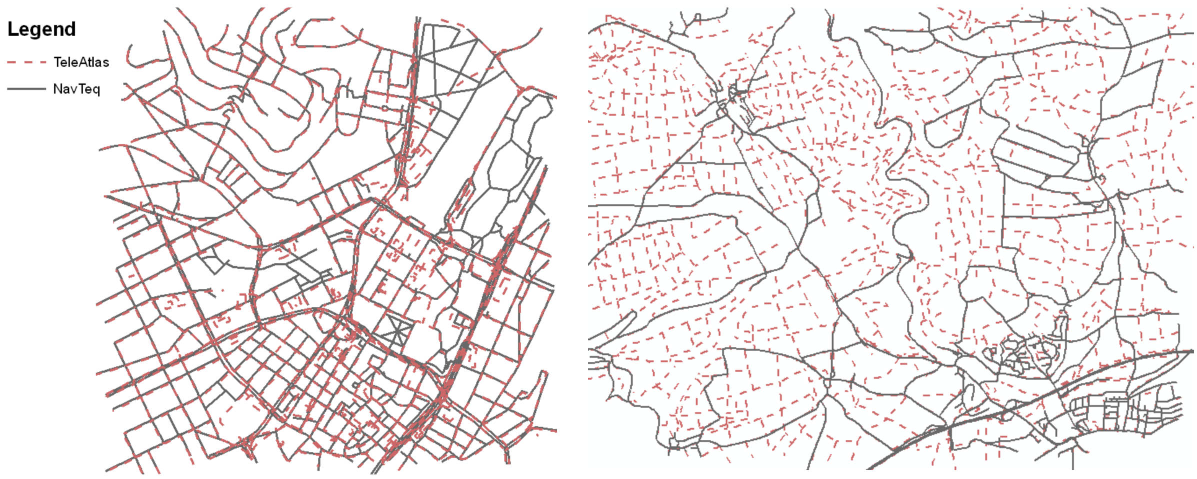


Figure 1: Overview of test area I (left) and test area II (right)

Manual Matching

To consider the topological differences between the two datasets, we extended the model presented in (Walter & Frisch, 1999) in order that not only matchings between edges but also between edges and nodes are possible. For example, node n_1 in Figure 2 (left) is matched to edge e_1 (Relation *P:1*). In **Figure 2** (right) the node n_1 is matched to four edges e_1, e_2, e_3, e_4 (Relation *P:n*).

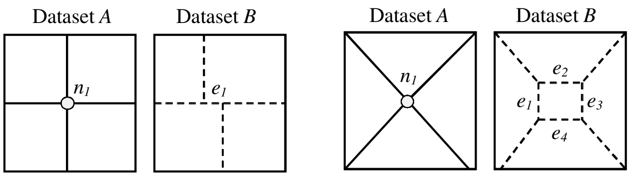


Figure 2: Matching between one node and one edge (*P:1*) and between one node and four edges (*P:n*)

The matching in our study is performed manually with a software tool developed with VBA and ArcGIS. **Table 2** summarizes the results of the manual matching and indicates that there are many differences between the two datasets even though they are representing the same objects of the landscape and they are collected in the same data model.

Table 2: Result of manual matching

Relation (NT : TA)	**Test area I (urban area)**			**Test area II (rural area)**		
	Matching	**NavTeq (NT)**	**Tele Atlas (TA)**	**Matching**	**NavTeq (NT)**	**Tele Atlas (TA)**
1:1	476 (53.7%)	476	476	353 (60.1 %)	353	353
n:1	69 (7.8%)	142	69	20 (3.4 %)	44	20
1:n	163 (18.4 %)	163	387	129 (22.2 %)	129	361
n:m	126 (14.0 %)	308	365	64 (10.9 %)	142	203
1:P	18 (2.0 %)	18	-	6 (1.0 %)	6	-
n:P	3 (0.3 %)	10	-	-	-	-
P:1	28 (2.3 %)	-	28	15 (2.6 %)	-	15
P:n	3 (0.3 %)	-	6	-	-	-
*1:**	-	176	-	-	16	-
**:1*	-	-	382	-	-	1039
Total	**886**	**1293**	**1713**	**587**	**690**	**1991**

Form Recognition

After the manual matching of the edges, the form of the edges of each matching pair is classified into different classes in order to derive automatically the matching of the nodes. The form is classified into eight basic classes according to the topology (see **Figure 3**). The class “Mix” is a combination of two or more basic classes.

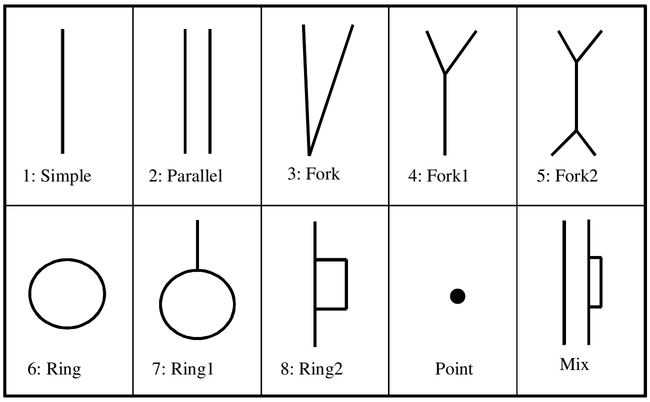


Figure 3: Different form classes

To identify the form, a network algorithm is implemented. For each dataset all edges of each matching pair are converted into a network. The node degree of each node in each network is calculated. According to this degree, the nodes are classified as following:

- *start* or *end node*: degree = 1
- *intermediate point*: degree = 2
- *intermediate node*: degree > 2

Depending on the node types, all networks are separated into several parts:

- *begin*: part from *start node* to *intermediate node*
- *middle*: part from one *intermediate node* to another *intermediate node*
- *end*: part from *intermediate node* to *end node*
- *whole*: part from *start node* to *end node*

The network in **Figure 4** of dataset *A* (solid line) consists of five parts and includes one "Ring2" (one begin, two middle and one end part) and one "Simple" (one whole part). Therefore, the form class is "Mix". The form class of dataset *B* (dashed line) is "Simple", because the corresponding network includes only one whole part.

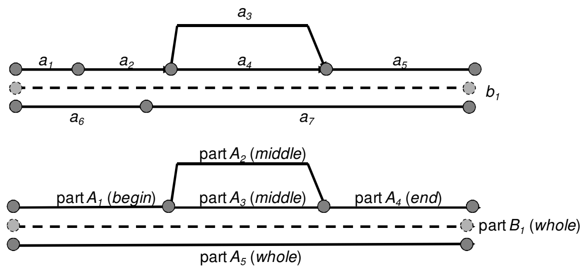


Figure 4: Parts of network

Node Matching

After the recognition of the form, the start and end nodes of the networks can be matched automatically (see **Figure 5**). Since a node may be included in more than one matching pair, the relations of node matching are also divided into *1:1*, *1:n*, *n:1* and *n:m*.

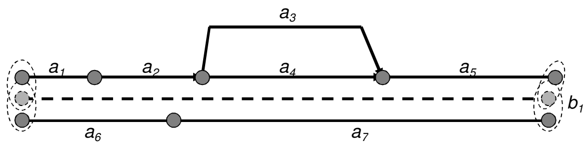


Figure 5: Automatic node matching

QUALITY INSPECTION

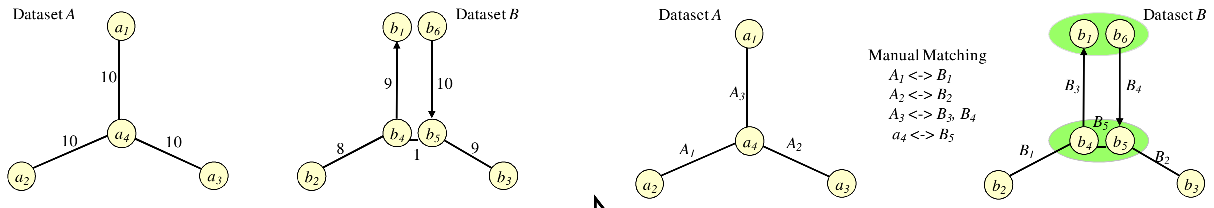
In the following, quality inspection methods based on similarity measurement on the levels of dataset and matching pair are presented. High similarity is an indicator for high relative quality.

Quality Inspection on Dataset Level

The quality measure on dataset level is calculated by comparing the adjacency matrices of the datasets. The precondition for this task is that the two adjacency matrices have the same dimensions and that the rows and columns are representing the same objects. Since there are many differences in the geometries of the two datasets, we introduce complex features in order to derive adjacency matrices that are comparable.

Geometric Similarity

Figure 6 shows an example of the building of complex features. The dimension of the adjacency matrix in dataset *A* is 4×4 and in dataset *B* 6×6. Based on the result of the manual matching of the edges and the following automatic node matching, the nodes (b_4, b_5) and (b_1, b_6) in dataset *B* are combined to complex nodes and the edges B_3 and B_4 to a complex edge. Nodes (a_2, a_3) and (b_2, b_3) are complex nodes with only one simple node. For complex edges the average length of their edges is used as length in the adjacency matrix. The two adjacency matrices become comparable on the complex feature level.



	a_1	a_2	a_3	a_4
a_1	-	-	-	10
a_2	-	-	-	10
a_3	-	-	-	10
a_4	10	10	10	-

	b_1	b_2	b_3	b_4	b_5	b_6
b_1	-	-	-	-	-	-
b_2	-	-	-	8	-	-
b_3	-	-	-	-	9	-
b_4	9	8	-	-	1	-
b_5	-	-	9	1	-	-
b_6	-	-	-	-	10	-

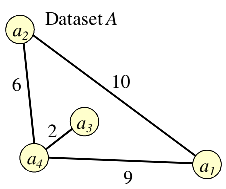
	a_1	a_2	a_3	a_4
a_1	-	-	-	10
a_2	-	-	-	10
a_3	-	-	-	10
a_4	10	10	10	-

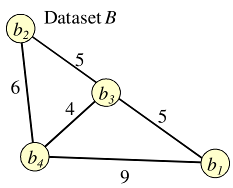
Node Matching
a_1 <-> b_1, b_6
a_2 <-> b_2
a_3 <-> b_3
a_4 <-> b_3, b_5

	b_1b_6	b_2	b_3	b_4b_5
b_1b_6	-	-	-	9.5
b_2	-	-	-	8
b_3	-	-	-	9
b_4b_5	9.5	8	9	-

Figure 6: Building of complex features

Because of topologic differences between the two datasets, there can be cells in the adjacency matrices which have a value in one of the matrices but not in the other, which means that there is no direct connection between these two nodes (see **Figure 7**). In that case we calculate the shortest path using the Floyd algorithm (Bondy & Murty, 2008) between these two nodes and use the length of this path as value in the adjacency matrix. After performing the Floyd algorithm, all elements in the adjacency matrices are comparable.





	a_1	a_2	a_3	a_4
a_1	-	10	-	9
a_2	10	-	-	6
a_3	-	-	-	2
a_4	9	6	2	-

	b_1	b_2	b_3	b_4
b_1	-	-	5	9
b_2	-	-	5	6
b_3	5	5	-	4
b_4	9	6	4	-

Floyd

	a_1	a_2	a_3	a_4
a_1	-	10	11	9
a_2	10	-	8	6
a_3	11	8	-	2
a_4	9	6	2	-

	b_1	b_2	b_3	b_4
b_1	-	10	5	9
b_2	10	-	5	6
b_3	5	5	-	4
b_4	9	6	4	-

Figure 7: Eliminating topologic differences with the Floyd algorithm

The maximum and average differences of the cells of the two adjacency matrices can be used as global quality measures. Absolute differences (meter) and relative differences (percent) are calculated. The result of the quality measurement on dataset level is shown in **Table 3**. By using Floyd algorithm 74 topologic differences in test area I and 38 in test area II are eliminated. Therefore, the average differences in both test areas and the maximal difference in test area I are reduced. The maximal difference in test area II remains due to large differences of the length of matched elements at the boundary of the test area. An average absolute deviation of 12.1 meters after Floyd can be observed in test area I and 27.7 meters in test area II. However, the average relative difference (14.8% in test area I and 12.5% in test area II) is almost the same.

Table 3: Result of geometric similarity inspection on the dataset level

	Test area I (urban area)		**Test area II (rural area)**	
	Before Floyd	After Floyd	Before Floyd	After Floyd
Number of Elements	1726	1726 (74)	1142	1142 (38)
Avg. Difference	13.8 m (16.0%)	12.1 m (14.8%)	29.7 m (13.9%)	27.7 m (12.5%)
Max. Difference	365 m (89.0%)	237 m (89.0%)	1397 m (89.1%)	1397 m (89.1%)

Topologic Similarity

The topologic similarity is calculated by evaluating the eccentricities of nodes. The eccentricity *E(x)* of a node *x* is the maximal distance (cost) of shortest paths from the node *x* to all other nodes in a network. Eccentricities of all nodes can be represented as a vector of eccentricity (N×1), where *N* is the amount of nodes in the network. The following eccentricity vectors are obtained for the networks of dataset *A* and dataset *B* in **Figure 7**.

$$\vec{E}\begin{pmatrix} a_1 \\ a_2 \\ a_3 \\ a_4 \end{pmatrix} = \begin{pmatrix} 11 \\ 10 \\ 11 \\ 9 \end{pmatrix} \text{and } \vec{E}\begin{pmatrix} b_1 \\ b_2 \\ b_3 \\ b_4 \end{pmatrix} = \begin{pmatrix} 10 \\ 10 \\ 5 \\ 9 \end{pmatrix}$$

The topologic similarity is defined as the correlation between the eccentricity vectors. The correlation is calculated as following, where *N* is the amount of elements, x_1 are the elements of vector *A* and x_2 of vector *B* (Wong & Lee, 2005):

$$r = \frac{N\sum(x_1x_2) - (\sum x_1)(\sum x_2)}{\sqrt{[N\sum x_1^2 - (\sum x_1)^2][N\sum x_2^2 - (\sum x_2)^2]}}$$

If the number of nodes in the datasets is not equal, the correlation between the eccentricity vectors cannot be calculated. Therefore, the eccentricity is calculated on the complex feature level. The result of the topologic similarity measure is summarized in **Table 4**. The topologic similarity in the urban area (test area I) is higher than that in the rural area (test area II). From **Figure 1** we can see that there are more differences in test area II than in test area I.

Table 4: Result of topologic similarity measure

	Test area I (urban area)	**Test area II (rural area)**
Correlation	0.928	0.840

Completeness

To measure the completeness, large datasets can be subdivided into small areas (e.g. 2×2 km) and the total length of the edges in each small area can be compared (Haklay, 2008). The completeness is calculated as following:

$$Completness_A(\%) = \frac{\sum Length_A}{\sum Length_A \cup \sum Length_B} * 100$$

Since the two test areas are not large, a further division is not necessary. In this paper the completeness is calculated on the level of simple feature and complex feature. **Table 5** summarizes the results of completeness inspection. The completeness of the two datasets in test area I is almost the same. However, the completeness of *TeleAtlas* in test area II is higher than that of *NavTeq* (see **Figure 1**).

Table 5: Result of completeness inspection

	Test area I (urban area)			Test area II (rural area)		
	NavTeq	TeleAtlas	Deviation	NavTeq	TeleAtlas	Deviation
Simple Feature	87.1%	88.7%	1.6%	42.7%	99.4%	56.7%
Complex Feature	85.2%	86.7%	1.5%	42.2%	98.5%	56.3%

Quality Inspection on Matching Pair Level

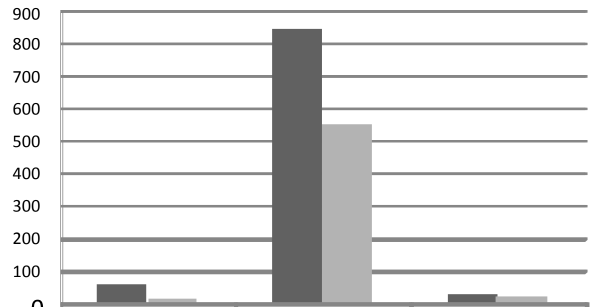
On the matching pair level the similarity of geometric modeling, geometric similarity and similarity of attributes are measured.

Similarity of Geometric Modeling

The similarity of geometric modeling is calculated based on the result of the form recognition. The different parts of the network are assigned with different weights: whole = 1; begin, middle and end = $\frac{1}{\text{number of different parts}}$. For example, the class "Fork1" (see Figure 3) consists of two different types of parts (one part begin, two parts end). Therefore, each part of "Fork1" obtains a weight $\frac{1}{2}$ and the sum of weights is $\frac{3}{2}$. The similarity of geometric modeling is equivalent to the ratio of the sum of weights in dataset *A* and dataset *B*:

$$Similariy_{Form} = \frac{\sum Weights_A}{\sum Weights_B}$$

A value nearby 1 indicates high similarity. If the value of similarity is greater than 1, the geometric modeling in dataset *A* of this matching pair is more complex than in dataset *B*. Otherwise the geometric modeling in dataset *B* is more complex than in dataset *A*. The result of the geometric modeling evaluation is summed up in **Figure 8**. More than 90% of all matching pairs are modeled similar, whereas only about 54% of all matching pairs in test area I and 60% in test area II are matched with 1:1 Relations (see **Table 2**). Since the two datasets are collected in same data model, the similarity evaluation provides a more plausible result for the geometric modeling evaluation than the evaluation of the matching relations.



	<1	=1	>1
Area I	54 (5.8 %)	848 (91.8 %)	22 (2.4 %)
Area II	7 (1.2 %)	555 (95.5 %)	19 (3.3 %)

Figure 8: Result of similarity of geometric modeling

Geometric Similarity

Different approaches are suggested in the literature for the measurement of the geometric similarity of linear features. For example, Goodchild and Hunter (1997) developed a method for evaluating the geometric similarity using buffers. The geometric similarity equates to the ratio between the length of linear feature inside of the buffer and the length outside of the buffer. Hausdorff distance (Deng et al., 2007; Hangouet, 1995) and Fréchet distance (Devogele, 2002) can also be applied to calculate the geometric similarity.

The geometric similarity in our approach is calculated with the Hausdorff distance. The result of the evaluation of the geometric similarity of all matching pairs can be seen in **Figure 9**. The average Hausdorff distance is smaller than 10 meters and meets the quality statement of the map suppliers. However, the large differences of edges at the boundary of test area II cause some large deviations.

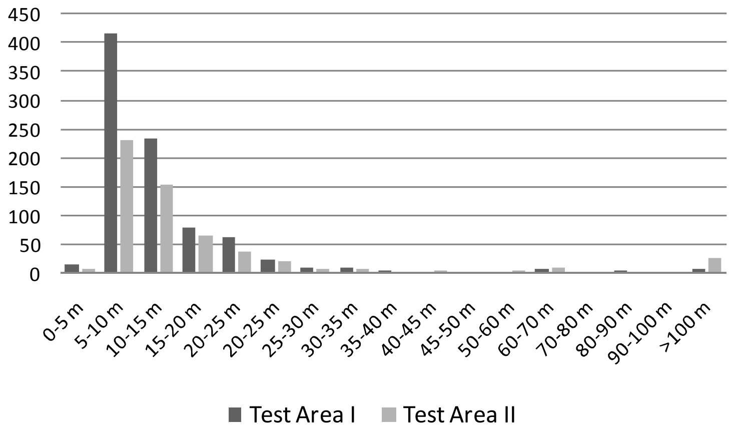


Figure 9: Result of geometric similarity

Similarity of Attributes

Depending on the attribute scale in each dataset, different methods are applied to calculate the similarity of attribute (see **Table 6**).

Table 6: Methods for attribute similarity inspection

Attribute Scale	Method	Attribute Example
Scales with identical finite range	Direct Comparison	Toll Road Ferry Ownership
Scales with identical infinite range	Direct Comparison	Street Name
Scales with different range	Confusion matrix	Functional Road Class

The similarity of attributes which are direct comparable is 1 if the attributes are identical, otherwise 0. A confusion matrix is used to calculate the similarities of attributes which have different ranges. For example, an attribute has k categories in dataset A and m categories in dataset B. The confusion matrix is calculated (see **Table 7**), where O_{ij} is the number of attribute matching between category i in dataset A and category j in dataset B, O_{i+} sum of row i, O_{i+} sum of column j and N sum of all elements in the matrix.

Table 7: Confusion matrix for calculating the similarity of attributes

	Dataset *B*				
Dataset *A*	***1***	***2***	**…**	***m***	***Sum*** $\sum$
1	O_{11}	O_{12}	…	O_{1m}	O_{1+}
2	O_{21}	O_{22}	…	O_{2m}	O_{2+}
…	…	…	…	…	…
k	O_{k1}	O_{k2}	…	O_{km}	O_{k+}
Sum $\sum$	O_{+1}	O_{+2}	…	O_{+m}	*N*

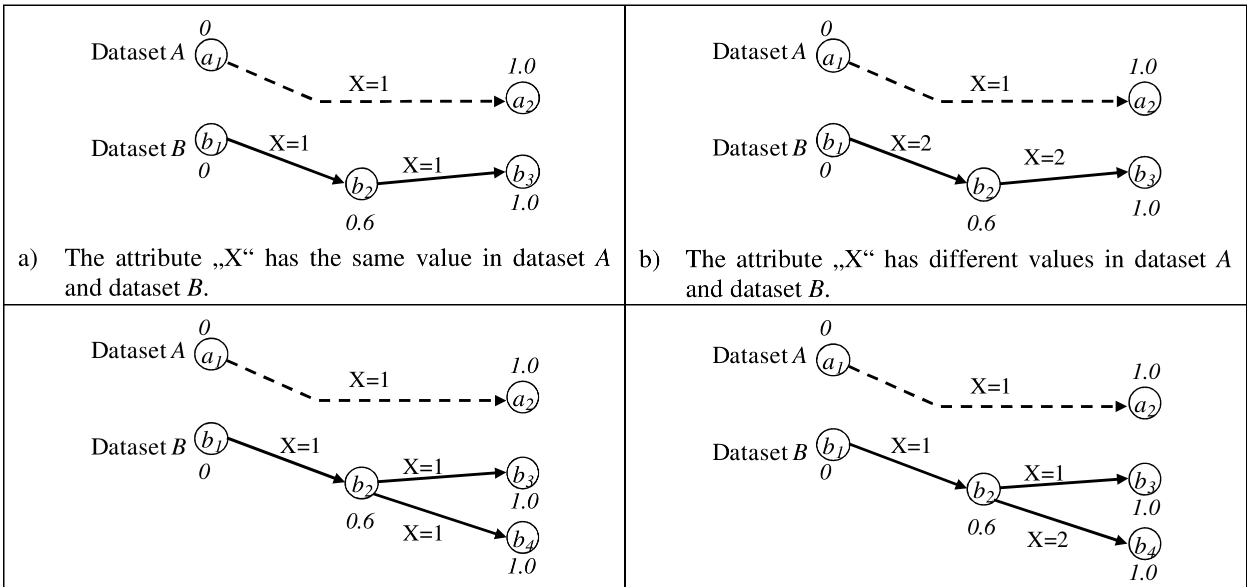
For an incomparable attribute, the similarity of attribute matching $s_{i,j}$ between category *i* in dataset *A* and category *j* in dataset *B* is calculated as following:

$$s_{i,j} = \frac{O_{i,j}}{O_{+j}}$$

If one attribute has different values (because the geometry of the corresponding matching pair consists of several parts), then the similarity is calculated as following, where $d_{i,j}$ is the distance along the matching pair of a segment with category *i* in dataset *A* and category *j* in dataset *B* and $s_{i,j}$ is the similarity of the attribute:

$$Similarity_{Att} = \frac{\sum(d_{i,j} * s_{i,j})}{\sum(d_{i,j})}$$

Figure 10 shows four examples of similarity measure for attribute "X" (two categories in dataset *A* and dataset *B*). The distance along the matching pair is also shown.



a) The attribute „X“ has the same value in dataset *A* and dataset *B*.

b) The attribute „X“ has different values in dataset *A* and dataset *B*.

c) The attribute „X“ has the same value in dataset *A* and dataset *B*.

d) The attribute „X“ has different values in dataset *A* and dataset *B*.

Figure 10: Examples of attribute similarity measurement

a) $$Similarity_X = \frac{(0.6 * s_{1,1} + 0.4 * s_{1,1})}{(0.6 + 0.4)} = \frac{(0.6 * 1 + 0.4 * 1)}{(0.6 + 0.4)} = 1.0$$

b) $$Similarity_X = \frac{(0.6 * s_{1,2} + 0.4 * s_{1,2})}{(0.6 + 0.4)} = \frac{(0.6 * 0 + 0.4 * 0)}{(0.6 + 0.4)} = 0.0$$

c) $$Similarity_X = \frac{(0.6 * s_{1,1} + 0.4 * s_{1,1} + 0.4 * s_{1,1})}{(0.6 + 0.4 + 0.4)} = \frac{(0.6 * 1 + 0.4 * 1 + 0.4 * 1)}{(0.6 + 0.4 + 0.4)} = 1.0$$

d) $$Similarity_X = \frac{(0.6 * s_{1,1}) + (0.4 * s_{1,1}) + (0.4 * s_{1,2})}{(0.6 + 0.4 + 0.4)} = \frac{(0.6 * 1) + (0.4 * 1) + (0.4 * 0)}{(0.6 + 0.4 + 0.4)} \approx 0.71$$

Table 8 shows the result of similarity measure for attribute scales with identical finite range (binary classification). The similarity is divided into three categories according to its value. The attributes *toll road* and *ferry* are 100% identical in both test areas. The average similarity of *ownership* in both areas is also very high (>0.99).

Table 8: Result of similarity measure for attribute scales with identical finite range (binary classification)

	Test area I (urban area)			**Test area II (rural area)**		
	Toll Road	*Ferry*	*Ownership*	*Toll Road*	*Ferry*	*Ownership*
Similarity=1	835 (100%)	835 (100%)	832 (99.7%)	566 (100%)	566 (100%)	561(99.1%)
0<Similarity<1	0 (0%)	0 (0%)	2 (0.2%)	0 (0%)	0 (0%)	1 (0.2%)
Similarity=0	0 (0%)	0 (0%)	1 (0.1%)	0 (0%)	0 (0%)	4 (0.7%)
Avg. Similarity	1.0	1.0	0.998	1.0	1.0	0.991

Table 9 summarizes the results of similarity measure for attribute scales with identical infinite range (street name). About 1.1% of the matching pairs in test area I and 5.5% in test area II do not have a street name. The average similarity of *street name* in test area I is higher than test area II.

Table 9: Result of similarity measure for attributes scale with identical infinite range (street name)

	Test area I (urban area)	**Test area II (rural area)**
Similarity=1	688 (82.4%)	358 (63.3%)
0<Similarity<1	25 (3.0%)	29 (5.1%)
Similarity=0	113 (13.5%)	148 (26.1%)
Attribute is not available in both datasets	9 (1.1%)	31 (5.5%)
Avg. Similarity	0.852	0.697

Table 10 indicates the result of the similarity measure for attribute scale with different range. The attribute "functional road class" has five categories in *NavTeq* and nine categories in *TeleAtlas*. The similarity of *functional road class* in test area II is higher than test area I, because there are fewer categories in the rural area (3 categories in *NavTeq* and 6 categories in *TeleAtlas*) than in the urban area (4 categories in *NavTeq* and 8 categories in *TeleAtlas*).

Table 10: Result of similarity measure for attribute scale with different range (functional road class)

	Test area I (urban area)	**Test area II (rural area)**
Avg. Similarity	0.774	0.855

DISCUSSION

The matching of the edges could also be done with an automatic matching approach. However, automatic matching can provide in some cases inaccurate matching results. These may cause inaccuracies in the quality measurement. In order to avoid this problem and to concentrate on the tasks after matching, we matched the edges manually. Many automatic matching approaches have been proposed in the last years. A matching rate between 90 to 95% can be typically achieved if the datasets are captured in similar scale (e.g. ATKIS, GDF). Therefore, we will add an automatic matching in the future in order to make this approach fully automatic.

The extended "Buffer growing" model is the basis for managing and interpreting the different kinds of modeling in the two datasets:

- Differences can be often found at intersections. Complex intersections ate often collected differently by different operators.
- Short roads with a length less than ten meters are collected either with nodes or with edges depending on

the operator and the accuracy of data collection.
- An undefined area or place may also be captured differently from operator to operator.
- In some cases physical barriers on the roads may be removed or added dynamically. This leads to different modeling of the same road at different time. For example, a road may be collected as a single or multiple carriageways depending on the time of collection.

In our approach we match the edges on the geometric level and calculate complex features according to the matching result. On the complex feature level the noise of the length caused by the differences of geometric modeling is reduced. This is a bottom-up approach.

A top-down approach is also possible: At first, complex features can be calculated by generalization in order that the differences in the datasets are removed or at least minimized. Then, the matching can be performed on the complex feature level and transformed afterwards to the geometric level. For datasets with different LODs, complex features must not necessarily be calculated. The matching can be done initially on the abstract levels and transformed gradually to the more detailed levels.

On the complex feature level, the adjacency matrices have the same dimensions and are comparable. The calculated maximum and average differences can be used as a global geometric accuracy measure. The deviations between the adjacency matrices before and after Floyd algorithm can be used to identify the differences of the topologic modeling in the datasets.

Based on the result of the evaluation of the geometric modeling, a map conflation (integration) can be realized. Matching pairs which are modeled similar can be conflated by calculating the middle line. Matching pairs with different geometric modeling can be grouped into different clusters and fused by a transformation of the clusters.

From the results of attribute similarity measure we can see that attribute scales with identical finite ranges have a higher similarity, because this attribute type can be simply differentiated during the data collection. Scales with infinite ranges have lower similarity due to the uncertainty during the data collection. For example:
- Attribute has multiple values in the reality, e.g. different street names "Königstraße" and "Rotebühlstraße" for same street. The street name is collected differently by different companies.
- Attribute is collected with different semantic by different companies, e.g. "Leonhardstraße" and "Leonhardplatz" for same street (straße=street; platz=place).

SUMMARY AND FUTURE WORK

In this paper we introduced an approach for data quality inspection based on map matching and data integration. In the first part we presented the matching model and the approach for form recognition and automatic node matching. In the second part of the paper we described a hierarchical quality inspection based on the result of matching.

Two commercial datasets (GDF/NavTeq and GDF/TeleAtlas) were compared in the research. After comparing the results of the evaluation on the different levels in the two test areas, we can conclude that the geometric accuracy, completeness and topologic accuracy in the urban area is higher than in the rural area. The reasons for the deviations in the two test areas need to be further investigated. The approach can also be used for data inspection and integration of commercial datasets and datasets from user generated content (e.g. OpenStreetMap). This can help to describe the quality characteristics of user generated content.

User interfaces will be developed in the future for a semiautomatic quality inspection. The results of the quality inspection on the level of dataset, matching pair and attribute can be categorized into different classes and visualized graphically with different colors (e.g. a traffic light model with red, yellow and green). The user interface can enable the operator to navigate into different level of details and analyze the deviations with selected or filtered information.

REFERENCES

Becker, C., Ziems, M., Büschenfeld, T., Heipke, C., Müller, S., Ostermann, J. and Pahl, M. 2008: Multi-hierarchical Quality Assessment of Geo-Spatial Data. In: Proceedings of the International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Beijing, XXXVII/B2, 779-786.

Bondy, J. A. and Murty, U. S. R. 2008: Graph theory. Springer, New York.

Casado, M. L. 2006: Some Basic Mathematical Constraints for the Geometric Conflation Problem. In: Proceedings of the 7th International Symposium on Spatial Accuracy Assessment in Natural Resources and

Environmental Sciences, 264-274.
Caspary, W. and Joos, G. 1998: Quality Criteria and Control for GIS Databases. In: Proceedings of the IAG SC4 Symposium, Eisenstadt, Austria, on CDROM.
Cobb, M. A., Chung, M. J., Foley, H., Petry, F. E., Shaw, K. B. and Miller, H. V. 1998: A Rule-based Approach for the Conflation of Attributed Vector Data. Geoinformatica, 2/1, 7-35.
Deng, M., Li, Z. L. and Chen, Z. L. 2007: Extended Hausdorff distance for spatial objects in GIS. International Journal of Geographical Information Science, 21/4, 459-475.
Devogele, T. 2002: A new Merging Process for Data Integration Based on the Discrete Fréchet Distance. In: Proceedings of the ISPRS Commission IV Symposium: Geospatial Theory, Processing and Applications, Ottawa, Canada, on CDROM.
Deretsky, Z. and Rdony, U. 1993: Automatic Conflation of Digital Maps. In: Proceedings of IEEE - IEE Vehicle Navigation & Information Systems Conference, Ottawa, A27-A29.
Edwards, D. and Simpson, J. 2002: Integration and Access of Multi-source Vector Data. In: Proceedings of the Joint International Symposium of Geospatial Theory, Processing and Application, Ottawa, Canada, on CDROM.
Frank, A. U. 2008: Data Quality-What Can an Ontological Analysis Contribute. In: Proceedings of the 8th International Symposium on Spatial Accuracy Assessment in Natural Resources and Environmental Sciences, Shanghai, China, 393-397.
Franz, B. 2008: Plausibilitätsprüfung von Karten für Navigationssysteme auf Basis aufgezeichneter Strecken. Diplomarbeit, Fachhochschul-Masterstudiengang SOFTWARE ENGINEERING, Hagenberg, Austria.
Gerke, M., Butenuth, M., Heipke, C. and Willrich, F. 2004: Graph-supported verification of road databases. ISPRS Journal of Photogrammetry and Remote Sensing, 58/3-4, 152-165.
Goodchild, M. F. and Hunter, G. J. 1997: A simple positional accuracy measure for linear features. International Journal of Geographical Information Science, 11/3, 299-306.
Goodchild, M. F. 2007: Citizens as sensors: the world of volunteered geography. GeoJournal, 69/4, 211-221.
Haklay, M. 2008: How good is OpenStreetMap information? A comparative study of OpenStreetMap and Ordnance Survey datasets for London and the rest of England. http://www.ucl.ac.uk/~ucfamha/OSM%20data%20analysis%20070808_web.pdf (last visit: 1st March 2010).
Hangouet, J. F. 1995: Computation of the Hausdorff Distance between Plane Vector Polylines. In: Proceedings of 10th International Symposium on Computer- Assisted Cartography, Bethesda, 4, 1-10.
Haunert, J.-H. 2005: Link based Conflation of Geographic Datasets. In: Proceedings of 8th ICA WORKSHOP on Generalisation and Multiple Representation, La Coruna, Spanien, on CDROM.
ISO14825 2004: GDF - Geographic Data Files - Version 4. Beuth Verlag, Berlin.
ISO19113 2002: Geographic information - Quality principles. Beuth Verlag, Berlin.
ISO19114 2003: Geographic information - Quality evaluation procedures. Beuth Verlag, Berlin.
May, I. 2002: Fortführung und Erweiterung von GDF (Geographic Data File) als Datengrundlage für Autonavigationssysteme. Dissertation, Universitätsverlag der TU Berlin.
Otto, H.-U. 2005: The ActMap Approach - Specifications of Incremental Map Updates for Advanced In-Vehicle Applications. http://ertico.webhouse.net/download/actmap_public_documents/2005-06%20ActMAP_approach_ITSinEurope.pdf (last visit: 1st March 2010).
Seo, S. and O'hara, C. G. 2009: Quality Assessment of Linear Data. International Journal of Geographical Information Science, 23/12, 1503-1525.
Visintainer, F., Darin, M., Flament, M., Durekovic, S., Otto, H.-U., Loewenau, J., Naumann, V., Andersson, H., Meliga, V., Thomas, B., Landwehr, M., Bimpas, M., Isaksson, P. and Haspel, U. 2008: Final requirements and strategies for map feedback. FeedMAP, D 2.2.
Volz, S. 2006: An iterative approach for matching multiple representations of street data. In: Proceedings of the JOINT ISPRS Workshop on Multiple Representations and Interoperability of Spatial Data, Hanover, Germany, 101-110.
Walter, V. and Fritsch, D. 1999: Matching Spatial Data Sets: a Statistical Approach. International Journal of Geographical Information Science, 13/5, 445–473.
Wong, D. W. S. and Lee, J. 2005: Statistical Analysis of Geographic Information with ArcView GIS and ArcGIS (Gebundene Ausgabe). John Wiley & Sons, Inc., Hoboken, New Jersey.
Zhang, M. and Meng, L. 2006: Implementation of a Generic Road-matching Approach for the Integration of Postal Data. In: Proceedings of 1st ICA Workshop on Geospatial Analysis and Modeling, Vienna, Austria, 141-154.