Integrating network structures of different geometric representations

S. Dalyot*[1], T. Dahinden[1], M. J. Schulze[1], J. Boljen[2] and M. Sester[1]

The availability of reliable vector geodata is increasing rapidly. However, there still exists a lack of appropriate tools and processes for integrated data management and analysis solutions that can handle the diversity of geodata, since structural, geometric and topological aspects affect their data modelling. This paper presents a process that is designed to handle not only existing geometric and topological differences but also structural differences associated with the interoperation and representation of 2D networks. While network structures (such as roads) are usually treated as areal objects in cadastre databases, they are also commonly treated as linear objects in topographic databases. Our integration method is designed to solve not only the positional conflicts in the geodata, but also the existing dissimilarities that are the result of different structural geometric representation primitives used. A localised geometric matching process is introduced for aligning these networks, in which distortions are monitored and quantified locally via sets of specifically selected observation constraints derived from the geometric structures. The aim is to assure that spatial consistency of the 2D geodata is maintained. The outcome presents a significant improvement of the initial state, suggesting a reliable solution for the problem of creating a homogenous unified geodata infrastructure with a statistically sound basis.

Keywords: Integration, Adjustment, Alignment, Geometry, Automation, GIS

This article has been peer reviewed and is from a recent FIG workshop/conference

Introduction

Inconsistencies in Geodatabases

Reliable vector geodata holds an important role in a variety of activities and applications, while the availability of such data is increasing rapidly. Geodata stored in vector databases results from extensive data collection and compilation, such that the Spatial Data Infrastructure (SDI) on which the databases are based on has also changed considerably in terms of its availability and quality.

Owing to the specification of standards, the aspiration is that geodata from different sources may be interchanged without major technical difficulties. Still, when overlapping data from different sources, it becomes apparent that it was collected under different conditions. These may differ in the specification of geometric primitives, data models, time of data acquisition, method of acquisition and observations, and quality assurance practice implemented to name a few. This leads to differences in semantics and geometry. Therefore, for an ingenuous user the data might seem to be incorrect, let alone an analysis using all data available might be difficult, if not at all possible.

[1]Institut für Kartographie und Geoinformatik, Leibniz Universität Hannover, Appelstraße 9a, 30167 Hannover, Germany
[2]Landesamt für Vermessung und Geoinformation Schleswig-Holstein, Mercatorstraße 1, 24106 Kiel, Germany

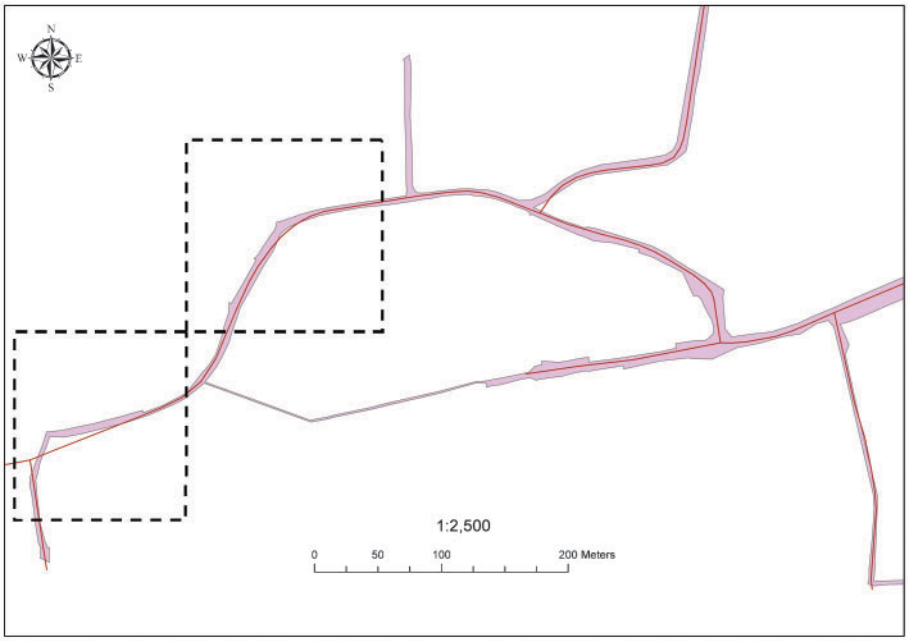
*Corresponding author, email sagi.dalyot@ikg.uni-hannover.de

This research deals with the special issue of the adaption of network objects for the creation of homogeneous geodata, mainly focusing on streets, paths, railways, and waterways. In large scale datasets these objects are typically represented with areal primitives, e.g. in cadastral data, whereas at smaller scales linear primitives are used, e.g. in topographic data, where the line segments composing the linear objects are representing an approximation of the middle axis position of the object in reality. The expectation of the user observing both representations is that the linear object should lie inside of the areal one. Obviously, as depicted in Fig. 1, this is not always the case. Due to existing geometric and topological differences (derived from various reasons stated in the previous paragraph), corresponding linear and areal objects do not share structural similarities and require some adjustment and integration in means of geometrical alignment. Still, such a process is not at all a simple procedure mainly due to the fact that no direct spatial geometric relations exist between the linear and the areal objects. A method that identifies and quantifies these relations (correspondences), together with the implementation of geometric alignment and adjustment that is based on these measures, is required for such cases.

Geodatabases integration solution strategies

A global solution strategy for the integration of networks stored in geodatabases should bring the

Received 27 June 2013; accepted 27 June 2013
DOI 10.1179/1752270613Y.0000000060



1 Problem definition: misalignment of geodata network features from different vector databases due to different representation and geometric inconsistencies: aim is that polyline feature class (red) will align to, e.g. exist within, polygonal one (pink) (rectangular areas are analysed later in the paper, see Fig. 9)

geodata into the same datum (thus eliminating some systematic inconsistency); e.g. geodata is adjusted on a high level, mostly followed by a global affine transformation. The outcome of such processes is geodata that is comparable on a global level: generally, the geodata fit together but still will show evident geometric discrepancies. To reduce such discrepancies, local solutions based on the shape and position of corresponding objects and the relation to other objects are used.

To facilitate a localised process, corresponding vertices existing in the networks can be identified and matched, followed by an adjustment procedure that makes use of a rubber sheeting. Still, even after this step, local discrepancies might exist. To eliminate such remaining discrepancies, a local geometric alignment process is used in this research. This process is based on the spatial geometrical relationships between point features (existing in one network dataset) and line features (existing in the other network dataset) of numerous objects [**2**].

Related work

Matching and alignment of networks

The matching and aligning of road datasets from different sources (Geodatabases) holds a vital role for processes, such as car navigation and topographical data, since the outcome has to maintain spatial consistency. Rosen and Saalfeld [**14**] addressed the problem of different geodatabases integration by concentrating on the matching of corresponding homologous point features that were identified and classified as centroids of road junctions, which were considered as the road networks foremost primitives. The two geodatabases were then adjusted using a rubber sheeting approach that used these identified uniform points existing in both geodatabases. This approach was later improved by several methods, such as [**4**], using a rule based matching approach regarding attributes, geometry and topology, and [**21**], enhancing matching to work not only with nodes, but also with segments and line edges.

Another matching approach is carried out by Gabay and Doytsher [**8**] which was able to detect the most common elements in both datasets and find elements which are not contained in both datasets. This allows the two datasets to have differences in geometry and topology where the algorithm is able to handle it. This approach is also carried out to be a two stage matching where line segments having corresponding end nodes are matched first. Based on this result unmatched line segments are further examined for their matching candidates. This matching approach was enhanced by Doytsher *et al.* [**6**] to work not on point based matching but to match whole line segments by looking at shape similarity, cumulative distance and similarity of emanating nodes in both end points.

The quantification of the extracted alignment results of road network datasets can also be used to improve other existing data in the vicinity that, to some geometrical or topological extent, has an affiliation to the road network datasets. Siriba [**15**] suggested extracting road features, which were implicitly given as gaps between polygons in a cadastral parcel dataset. The extracted polygons were then collapsed to polylines using a straight skeleton approach [**11**], [**12**] and [**13**], to be matched later to a given polyline road network (centrelines) of a topographical dataset, which had higher accuracy. The calculation of displacement vectors for the road middle axis was carried out via a rubber sheeting approach. These displacement vectors were then used to improve the accuracy of the cadastral dataset to enhance reliability of shape and position of land parcels. Still, this approach is also based on the fact that the aligned geodata holds the same geometrical structure, i.e. polylines.

A different approach for finding correspondences between different vector datasets was given by Dalyot

et al. [**5**]. This approach uses the correspondences between features by building perpendicular links between points existing in one dataset and their corresponding line segments existing in the other dataset. The datasets were then matched by using a Least Squares Adjustment (LSA) method aimed at minimising the corrections quantified by the perpendicular distances, considered as gaps and displacements, between corresponding points and line segments. The outcome of this localised matching process is two aligned datasets, which do not show significant discrepancies. However, this method only works with direct correspondences between line geometries. A solution to align objects with different geometric representations is presented in this paper.

A different line based matching approach was carried out in [**19**] and [**20**] on ATKIS (the German Topographic-Cartographic Information System) and GDF (commonly used format and model for data interchange of automotive navigation system) datasets. The algorithm searches for all potential correspondences of connected line segments using a buffer. As the list of correspondences is highly ambiguous, candidates are deleted which do not match certain relational parameters, such as topological information as well as feature based parameters like line-segment angles. However the resulting list is still ambiguous, the matching candidates are evaluated with a merit function to find a unique combination of matching pairs. This combinatorial problem is solved with an A* algorithm [**1**].

Another matching method for GDF and ATKIS data is presented by Volz [**18**]. The two datasets are coarsely adapted with a rubber sheeting transformation. Therefore warping nodes are identified in both datasets, showing a high level of correspondence and having at least four nodes with same length and angles. Then the translation vector field based on these nodes is analysed to remove strong outliers and is applied to the dataset. To ensure unique 1:1 matching partners, GDF and ATKIS are geometrically split up by projecting nodes from one dataset to the other and vice versa, if they are not suited near an existing point. Candidates for projection are identified using a buffer operation as well as looking at parallelism of their edges. The following iterative adjustment uses seed nodes which are identified the same way as with coarse transformation. The algorithm then starts from these nodes looking at neighbouring nodes to identify their matching partners by looking at similarity values such as distance and difference in the nodes' degree. Also edges are investigated and matched using similarity values (e.g. length difference or Hausdorff distance). These values are combined to receive a score value for similarity for nodes as well as one for edges. The determined score value as well as a case differentiation for the edges (a. both end nodes are matched, b. both are unmatched, c. only one is matched) a combined rule based and probability approach leads to determination of 1:1 correspondences. Unmatched nodes are adjusted in regard to their matched neighbours and therefore should find a matching partner in one of the next iterations starting again with identification of seed nodes.

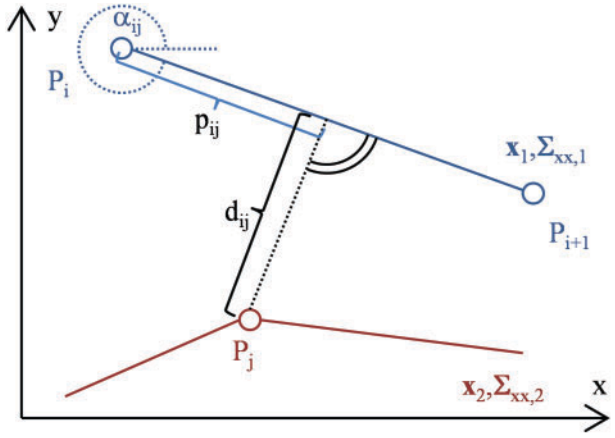
Integration of heterogeneous geospatial data into a common database was investigated by Butenuth et al. [**3**], emphasizing hydrographic water datasets, and geographic features, only. At first, corresponding objects were carried out by corresponding object class and geometry using symmetric differences (e.g. difference of union and intersect) between objects in the two databases as well as azimuth histograms. The alignment of identified similar objects is done using the Iterative Closest Point (ICP) algorithm as a first step and a dual interval alignment where corresponding vertices are identified using Euclidian distance as well as angle similarity. Each vertex can be aligned by calculating the shift to the corresponding vertex; however, this shift can be weighted if the object should not be fully aligned. This method was also tested with raster data by using a field boundary detection based on a segmentation followed by a network snake approach to align the identified field boundary to a vector dataset.

Networks alignment to raster datasets

Vector network datasets can also be improved by aligning them to raster datasets, derived from Airborne Laser Scans (ALS). Displacement errors occur due to different acquisition methods, geo-referencing methods and data structures. Network snakes are used to align the road network to the ALS data by minimising an energy term built on the summation of laser intensity, height information and other calculated parameters derived from the ALS data [**9**]. This approach is widened up to work with other raster datasets, such as stereo images, where energy is derived from radiometric and geometric features, e.g. height, in the images. Also, a self internal test is introduced in this approach to evaluate the processed results and allowing manual corrections, if necessary [**10**].

A different way of improving the position of GIS vector datasets by using aerial photogrammetry is presented in Filin and Doytsher [**7**]. This approach works on linear features, which are obtained in the vector dataset and in orthophotos. The algorithm to enhance accuracy of the vector dataset detects these linear features and matches them by looking at similar characteristics, such as shape similarities, cumulative distance and emanating nodes at both end points. Then the scene is divided up in small parts where for every part a local transformation is calculated and applied to remove the distortions between the two datasets.

Another method for matching features based on their characteristics is given by Vassilaki *et al.* [**16**]. This approach uses ICP for finding corresponding curves in a given network. The metric for defining the distance between corresponding curves is given by the maximum value of four characteristics (distance between first nodes, last nodes and centroid, as well as absolute distance of curve length). This matching approach was tested on road networks extracted from a topographic map and an IKONOS high resolution satellite image, giving the ability to align the satellite image to the topographic map. The centrelines of the road networks were derived from digitised road boundaries using a skeleton algorithm. The method was improved to also work with 3D free form linear features, allowing also the alignment of features with different dimensionality (2D to 3D) [**17**]. The algorithm was used on features digitised from topographic maps of different scales and time as well as radar satellite imagery to determine their alignment parameters.



2 Geometric dependency between two line features

Methodology

The modelling concept described here is designed to cope with the specific geometrical structural differences: areal (polygon) and linear (polyline) modelling, together with geometrical and topological aspects. It is aimed at overcoming problems, such as distribution and correspondence ambiguity. This is achieved by carrying out a localised adjustment model using feature and relational based rules in order to reduce alignment differences, which are monitored and quantified during the process.

Geometrical dependency

The sets of coordinates existing in two vector geodatabases are known and given: x_1 (reference) and x_2 (source), including their corresponding stochastic modelling matrices $\Sigma_{xx,1}$ and $\Sigma_{xx,2}$ respectively. The relation between the stochastic modelling matrices defines the weighting of each database in the process. To achieve geometric alignment, the perpendicular gaps d_{ij} (where i and j are two corresponding linear segments) between corresponding features from both databases are reduced to zero, and minimised coordinate corrections using a LSA method. The rectangular gaps are formed by calculating vertex to line segments distances, e.g. coupling-up pairs of corresponding vertices considered the nearest in the corresponding features, while maintaining the assumption of polyline-to-polygon alignment.

In general, when discussing the integration of road networks having different structures, the common case will show that a vertex in the polyline feature class 'should' have two corresponding vertices in the polygon feature class, e.g. to both polygonal road edges. When perfectly aligned, the polyline should have similar geometry to both polygonal faces. Thus, presuming that after the geometric alignment adaptation the polyline feature geometry has to lie within the polygonal one, e.g. lie on the polygon's middle axis. The algorithm is designed to improve the given coordinates in order to remove the gaps (regarded as contradictions) in the iterative LSA process.

Depicted in Fig. 2, the perpendiculars d_{ij} from a given source line vertex (point) P_j in x_2 to all candidates corresponding line segments in $\boldsymbol{x}_1$ are identified. The existing perpendicular gaps (d_{ij}) are calculated to each line segment P_iP_{i+1}. For vertex P_j there might exist zero to m such correspondences. The coordinates of both databases are adjusted in the way that the quadratic sum of these corrections is minimised under the conditions of eliminating the perpendiculars, while maintaining the condition that the source geometry (polyline) must lie within the reference geometry (polygon). The derogation along the line segment of x_1 is called p_{ij} and is defined together with d_{ij}, as depicted in equation (1), where $i=1,\ldots,n_d$, and, $j=1,\ldots,n_p$ (where n_d is the number of source corresponding vertices, and n_p is the number of reference corresponding vertices). The angular value α_{ij}, depicted in equation (2), describes the azimuth of the line segment P_iP_{i+1} and also implicitly the direction of the perpendicular d_{ij}.

$$\begin{aligned} P_{ij} &= (y_j - y_i)\sin\alpha_{ij} + (x_j - x_i)\cos\alpha_{ij} \\ d_{ij} &= (y_j - y_i)\sin\alpha_{ij} - (x_j - x_i)\cos\alpha_{ij} \end{aligned} \tag{1}$$

$$\alpha_{ij} = \arctan\left(\frac{y_{i+1} - y_i}{x_{i+1} - x_i}\right) \tag{2}$$

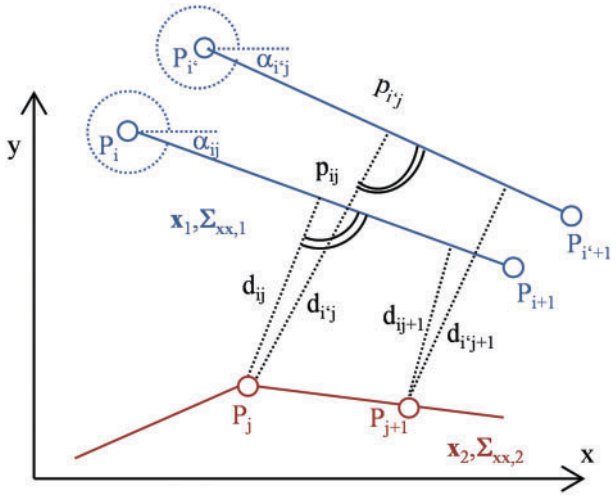
Still, one has to remember that the requirement here is to integrate two different structural geometries that represent road networks, e.g. polyline geometries onto polygon geometries. In this case, the desired outcome might suggest that polyline features will lie within the corresponding polygon ones. Thus, an assumption is made, in which all valid d_{ij}'s are chosen with respect to the fact that the value of the corresponding α_{ij}'s for a specific P_j are close to being parallel, as depicted in equation 3 (where $\alpha_{i'j}$ depicts the angle value of line $P_{i'}P_{i'+1}$ depicted in Fig. 3). Parallelism threshold α_c is chosen with respect to the datasets at hand and/or flexibility required in the output

$$\left|\alpha_{ij} - \alpha_{i'j}\right| \le \alpha_c \tag{3}$$

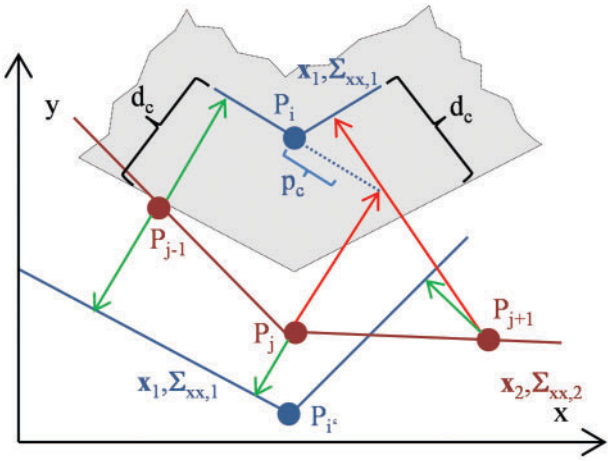
An example depicting such a desired geometric arrangement of two valid assignments is depicted in Fig. 3. The desirable outcome here will be for source vertex P_j to be shifted towards the area existing between the two line segments (in blue) – P_iP_{i+1} and $P_{i'}P_{i'+1}$, i.e. polygon edges. The alignment process will aim at minimizing the existing difference, considered in the LSA as residuals, between distances d_{ij} and $d_{i'j}$ (and also d_{ij+1} and $d_{i'j+1}$). Since both perpendicular distances d_{ij} and $d_{i'j}$ cannot be eliminated simultaneously while maintaining the abovementioned geometrical requirement, the condition is modified as depicted in equation 4, where an updated d_{ij}^* is calculated based on the difference value of both perpendiculars. The use of $\pm$ is made since the sign of values d_{ij} and $d_{i'j}$ is dependent on the direction of P_j in respect to both line-edges, and also on their absolute values. In general, where only two parallel candidates are found there exist two constellations (depicted in Fig. 4):

(i) source vertex falls inside (between) two candidate parallel line segments (as P_{j-1} in this example)
(ii) source vertex falls outside two candidate parallel line segments (as P_{j+1} in this example)

$$\begin{aligned} \Delta d &= \frac{\left|(d_{ij} + d_{i'j})\right|}{2} \\ d_{ij}^* &= d_{ij} \pm \Delta d \end{aligned} \tag{4}$$



3 Valid geometric dependency between source polyline segment (P_iP_{i+1}) and two parallel reference line segments of polygon feature



4 Geometrical assignments of source line segments (brown) in respect to reference line edges, and thresholds used: invalid (red) and valid (green) assignments; grey polygon depicts valid buffer area (d_c) of upper edge

The LSA minimises the quadratic sum of the coordinate corrections under the conditions stated in equation 4. A valid solution is obtained while assignments for the perpendiculars are chosen in respect of several criteria. At first, the width of the reference line segment (edge of polygon) for each perpendicular has to be defined via a buffer around every source line segment (of the polyline). It regards the maximum length of the perpendicular d_c (in respect to values of d_{ij}) and the length p_c (in respect to values of p_{ij}) that defines the maximum length for the extension of a reference line segment, where s_{ij} describes the total length of the reference line segment P_iP_{i+1} (depicted in Fig. 2). The definition of these two thresholds leads to the criteria equations depicted in equation (5)

$$\begin{aligned} -p_c \le p_{ij} \le s_{ij} + p_c \\ -d_c \le d_{ij} \le d_c \end{aligned} \tag{5}$$

All perpendiculars that are not valid in regard to equations (3) and (5) are not used in the adjustment process. The buffer for valid assignments is visualised in Fig. 4 (source line segments in brown and reference line segments in blue): red perpendiculars depict invalid assignments ($>|d_c|$ and/or $>|p_c|$) that are not used in the LSA. On the right (for vertex P_{j+1}) a value of d that is larger than the predefined d_c threshold, and on the middle (for vertex P_j) a correspondence that falls outside of the extension threshold value of p_c. The thresholds p_c and d_c have to be adjusted to match to a certain given geodata or a certain task.

Consequently, both coordinates of the databases are adjusted in the way that the quadratic sum of these corrections is minimised under the modified condition. It should be noted that the solution here is not limited to only one assignment per each source vertex; in case several candidate parallel line segments exist (as in the case of polygon corners, turns or junctions), i.e. more than two perpendicular residuals exist, the LSA process handles all as observations in order to achieve the desired geometrical alignment.

During the LSA P_j is moved toward a common position derived by all valid assignments to form a common polyline feature existing within the polygon feature. The scale of the adjustment depends on the number and distribution of valid perpendiculars, as well as the stochastic model of coordinates.

Least squares adjustment

Based on the functional model for the perpendiculars d_{ij} (equations (1) and (4)), an LSA is defined based on conditional observations. The conflict between the two datasets, denoted as w_x, is described by the value of the perpendiculars d_{ij}. Since both values for a single vertex P_j - d_{ij} and $d_{i'j}$ - are dependent observations, only one of the adjusted (in respect to equation (4)) is used in the LSA; the other is eliminated from the process. The functional dependency from the corrections of the coordinates, denoted as v_x, to the existing conflict w_x is described in matrix $\mathbf{B}_x$, depicted in equation (6), and therefore has to be linearised. A stochastic model is derived out of the covariance matrix of the point coordinates $\Sigma_{ll,x}$. The cofactor matrix of the point coordinates $Q_{ll,x}$ is then built by eliminating the variance factor σ_0^2 from $\Sigma_{ll,x}$:

$$\begin{aligned} &\mathbf{B_x} v_\mathbf{x} + w_\mathbf{x} = 0 \\ &\mathbf{B_x} = [\,\mathbf{B}_{\mathbf{x},1} \quad \mathbf{B}_{\mathbf{x},2}\,], v_\mathbf{x} = \begin{bmatrix} v_{\mathbf{x},1} \\ v_{\mathbf{x},2} \end{bmatrix} \\ &\sum_{\mathbf{ll,x}} = \sigma_0^2 \mathbf{Q_{ll,x}} \\ &\mathbf{Q_{ll,x}} = \begin{bmatrix} \mathbf{Q_{xx,1}} & 0 \\ 0 & \mathbf{Q_{xx,2}} \end{bmatrix} \end{aligned} \tag{6}$$

Each row of $\mathbf{B}_x$ corresponds to one perpendicular link between the two databases. The matrices and vectors can be split up in two parts containing only the functional dependency to coordinates of x_1 or x_2. The condition for the algorithm to work correctly is to set no correlations between the two datasets, therefore $Q_{ll,x}$ can be split up to a cofactor matrix for the coordinates of x_1, $Q_{xx,1}$ and a cofactor matrix for x_2, $Q_{xx,2}$. The linearised model derived from equations (1) and (4) describes the

following relationships for the coordinates, as depicted in equation (7)

$$\begin{aligned} & v_{x,i}\left(1-\frac{p_{ij}}{s_{ij}}\right)\sin\alpha_{ij}-v_{y,i}\left(1-\frac{p_{ij}}{s_{ij}}\right)\cos\alpha_{ij} \\ & +v_{x,i+1}\frac{p_{ij}}{s_{ij}}\sin\alpha_{ij}-v_{y,i+1}\left(1-\frac{p_{ij}}{s_{ij}}\right)\cos\alpha_{ij} \\ & -v_{x,j}\sin\alpha_{ij}+v_{y,j}\cos\alpha_{ij}+d_{ij}=0 \end{aligned} \tag{7}$$

The solution v_x, depicted in equation (8), is derived using the cofactor matrix of the conflicts Q_{ww}. The stochastic information of adjusted coordinates is contained in the adjusted observations cofactor matrix, and is determined by variance propagation

$$\begin{aligned} & Q_{ww,x}=\sum_{i=1}^{2}\mathbf{B}_{x,i}Q_{xx,i}\mathbf{B}_{x,i}^{T} \\ & \begin{bmatrix}\mathbf{v}_{x,1}\\ \mathbf{v}_{x,2}\end{bmatrix}=-\begin{bmatrix}Q_{xx,1}\mathbf{B}_{x,1}^{T}\\ Q_{xx,2}\mathbf{B}_{x,2}^{T}\end{bmatrix}Q_{ww,x}^{-1}w_x \end{aligned} \tag{8}$$

A multiple statistical testing procedure, depicted in equation (9), can be applied to the adjustment process. Still, the corresponding results must be interpreted carefully since some links accepted by the testing procedure violate logical geometrical relations, such that further development is necessary to surmount possible ambiguities of perpendicular links

$$|y_j|=\frac{w_{x,j}}{\sigma_o\left(q_{ww,x,j}\right)^{1/2}}\le y_{1-\alpha o} \tag{9}$$

Structural adaptation

When comparing different geodata existing in vector databases, most cases will show different number of vertices and line segments that are stored in each database (level-of-detail mostly differ). During an alignment process, the coarser feature structure might not be aligned completely to the finer feature structure. The quality of approximation for an object is given by the number and density of recorded points. A structure with a high density of points is always adjustable to a structure with a low density, but not vice versa; there is a bias towards the structure with the higher density of points, as it has more observations. This effect can be avoided by increasing the density with the interpolation (and introduction) of new points in the coarser database.

To eliminate as much as possible these geometric conflicts of databases having different vertex densities the geometric sub-division of the coarser database features is introduced and integrated in the process. The creation of new intermediate vertices in the process, considered as predetermined breaking points in the line segments that are required to be aligned, leads to a more homogenous geometrical adaptation of existing features. This enables higher flexibility of the more coarse features to be aligned to the finer ones.

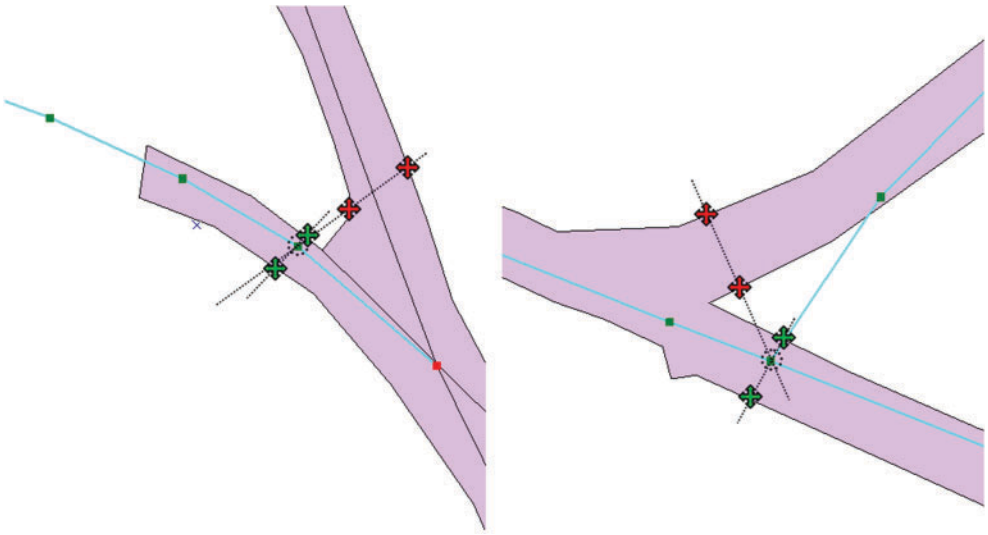
For this process every reference vertex of x_1 is projected via perpendicular assignments to the corresponding source line of x_2. A new vertex is interpolated and added to the line segment in x_2 if the perpendicular length d_{ji} and the distance to the next point on that line segment p_{ji} are valid, in respect of the equations depicted in equation (10). Therefore, the position of the new point on the corresponding line segment is given by the relation of p_{ji}/s_{ji}. This process is carried out as a preliminary process, when required, before the assignment of correspondences and LSA is carried out

$$\begin{aligned} & p_c\le p_{ji}\le s_{ji}-p_c \\ & -d_c\le d_{ji}\le d_c \end{aligned} \tag{10}$$

Additional assignments

As outlined earlier, the basic assignment requires that two close-to-parallel candidates exist in order to create a geometric correspondence between a vertex (source polyline) and the corresponding line segment candidate (reference polygon). Still, in order to facilitate more flexibility to the LSA, and to allow constellations that show otherwise, additional assignments are also used and implemented in this methodology.

1. Single correspondence:
   - 1.1 As shown in Fig. 4, vertex P_j has a single correspondence to only one of the polygon's edges, thus it has a single valid assignment, instead of two. Still, the aspiration is to be able to monitor such a case and still be able to use this single assignment in the LSA. Consequently, an additional constraint is used to detect such constellations: finding whether other polyline vertices in the vicinity (neighbouring vertices in the order of two) of that candidate vertex have existing assignments to that specific reference polygon. In case there are, the value of the average buffer is extracted, based on existing dual correspondences, later to be used as a single correspondence in the LSA process (e.g. value of Δd in equation (4) is extracted via existing neighbouring correspondences).
   - 1.2 As shown in Fig. 4, vertex P_{j+1} exist outside the reference polygon and has a single correspondence to only one of the polygon's edges, thus it has a single valid assignment - instead of two. Still, the aspiration is to be able to monitor such a case and still be able to use this single assignment in the LSA. Consequently, an additional constraint is used to detect such constellations: finding whether other neighbouring vertices of that candidate-vertex have existing assignments and also validating whether the direction of that source line segments (this vertex is a part of) exists in the parallelism threshold. In case both constrains are validated, the value of the required buffer is extracted based on the neighbouring dual correspondences. This value is used for the single correspondence in the LSA process (similar to 1·1 above).
2. Multiple correspondences:
   - 2.1 Multiple correspondences (more than two) may occur. Still, due to flexible threshold use and mainly in the vicinity of junctions, some correspondences might exist that can be considered as ambiguous ones. For example,



5 Reference polygon constellations with junction branches showing ambiguous valid assignments (red crosses) in respect to valid ones (green crosses) for single source vertex (circled)

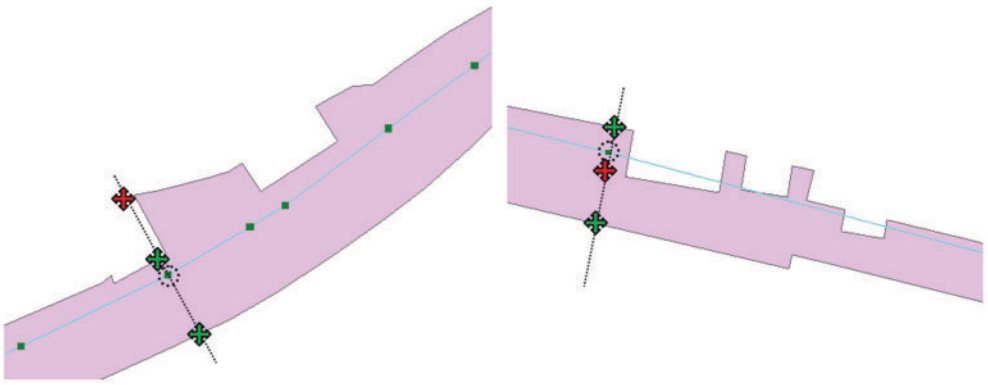
a source vertex that exists inside a branched road coming out from a junction that is pulled to another road branch; examples are depicted in Fig. 5. In case all correspondences are used in the LSA, the vertex will be pulled to some arbitrary position. Consequently, an additional constraint is used to detect such constellations, which detects the most appropriate branch to be used. This is achieved by deriving additional topologic values: general branch direction and containment - which branch direction (azimuth) is closest (most similar) to the source polyline and also contains the source segment. This results in monitoring incorrect assignments that are filtered out of the LSA process.

2.2 As shown in equation (5), the geometrical assignments make use of the line extension as a valid assignment ($-p_c$ or $s_{ij}+p_c$). Some reference line constellations will show several correspondences in the same direction due to sharp changes in the reference polygon geometry due to the use of, but not restricted to, the extension; examples are depicted in Fig. 6. Using all assignments will result in an incorrect outcome since several assignments try to pull a single source vertex in the same direction; since the vertex is already inside the polygon, this is an invalid solution. Also, as in the case of the circled vertex in the left image, it has three valid assignments (crosses), thus it is hard to derive what is the correct polygon buffer area needed to be used in the process, out of the three possible combinations. Consequently, an additional constraint is used to find such constellations: filtering out multiple assignments by validating that legitimate assignments pull the source vertex in opposite directions. Also, the value of the valid buffer size is derived from other neighbouring assignments. Such that several assignments are filtered out and only two are left to be used in the LSA.

Experimental results and discussion

Data

The algorithm was implemented and analysed on several datasets. The aim was to test it in integrating different vector databases representing network structures with different geometric representations. The geodata used represented different data sources used in Germany:



6 Reference polygon constellations showing ambiguous valid assignments (red crosses) aimed to be filtered for single source vertex (circled)

ALKIS (Amtliches Liegenschaftskataster-informationssystem, i.e. Authoritative Real Estate Cadastre Information System) - storing cadastral data with polygonal parcel features, including the networks; ATKIS (Amtliches Topographisch-Kartographisches Informationssystem, i.e. Authorative Topographic-Cartographic Information System), storing topographic data where network data is represented as polyline features. The scale of the ALKIS database (cadastral data) is of 1:5000, while the ATKIS database (topographic data) was designed for a scale of 1:25 000. As such, the polygonal data shows a higher level of detail and is also considered as more accurate. Therefore, the ALKIS data was used as the reference data with higher accuracy (higher weight in the LSA). Observed discrepancies of metres - and up to several dozens of metres, exist not only because of generalisation but also due to different data collection methods and level of detail.

Results

Figure 7 depicts a somewhat simple geometrical case, where a straight railway track is represented simultaneously as polygon (ALKIS) and polyline (ATKIS) features in both datasets. It should be noted that the ALKIS railway is presented by five different features in the database; this adds a certain ambiguity into the integration process since additional line segments (feature borders), which can be considered as redundant in the process presented here, are introduced. Two example areas are shown in Fig. 7 left and right, where (from top to bottom) the top row shows the initial geometric state, the middle row shows the outcome of the proposed methodology (final geometric alignment positioning), and the bottom row shows a zoomed-in area. Together with the input and output features, a vector field layer (i.e., v vector) is added (represented by arrows), which shows the displacement in magnitude and direction of each vertex from the source feature toward its position in the output.

Since ALKIS is used here as a reference, with higher weight value of 1:10 in the LSA process, only ATKIS vector fields exist. In terms of threshold values used, $(d_c,p_c)=(10, 3)$ m respectively. These values were chosen based on a rough estimation of the existing geometric displacements. Residual values of both datasets in the two iterations implemented are depicted in Table 1, showing that indeed both datasets converged, indicating geometric alignment.

It is observed that the entire polyline feature is moved with a constant trend from its initial position to its final one, which is on the middle axis of the polygon feature. It is also observed that the magnitude of each shift, represented by the vector field arrows, is derived from the buffer extent of the polygon feature for each specific area within the polygon, and that this shift is compliant with the reference geometry. This advantage is a direct result of using a local adjustment process, instead of a coarser global one. The example on the bottom right (boxed) shows the problem of more than two available parallel reference line segments discussed in the previous section. Still, the use of additional assignments here made it feasible to identify the consistent geometric changes in the reference feature, shifting the correspondence used for alignment from the wider polygon buffer to the narrower one.

Figure 8 depicts two more examples, where a junction is also involved in the integration process. As in the previous example, not only that the geometry of the railway track branches are aligned precisely, also in the vicinity of the junction the source polyline geometry is modified and changed. This creates a fitted curve where a straight line once was. Still, this geometric modification (boxed area on bottom left) raises a problem since it modifies the existing navigation system and rules: from a T-junction (blue line) to a right-hand merge (red line); thus, future work will entail addressing such tasks. The existence of curvatures, where once only straight segments existed, proves that indeed several assignments (and not merely a single one) were used in the LSA.

The same threshold values were used in both cases, which were chosen based on a rough estimation of the existing geometric displacements. Residual values of both datasets after the first iteration are depicted in Table 2, while the second iteration showed values close to zero, proving that both datasets converged, thus indicating geometric alignment.

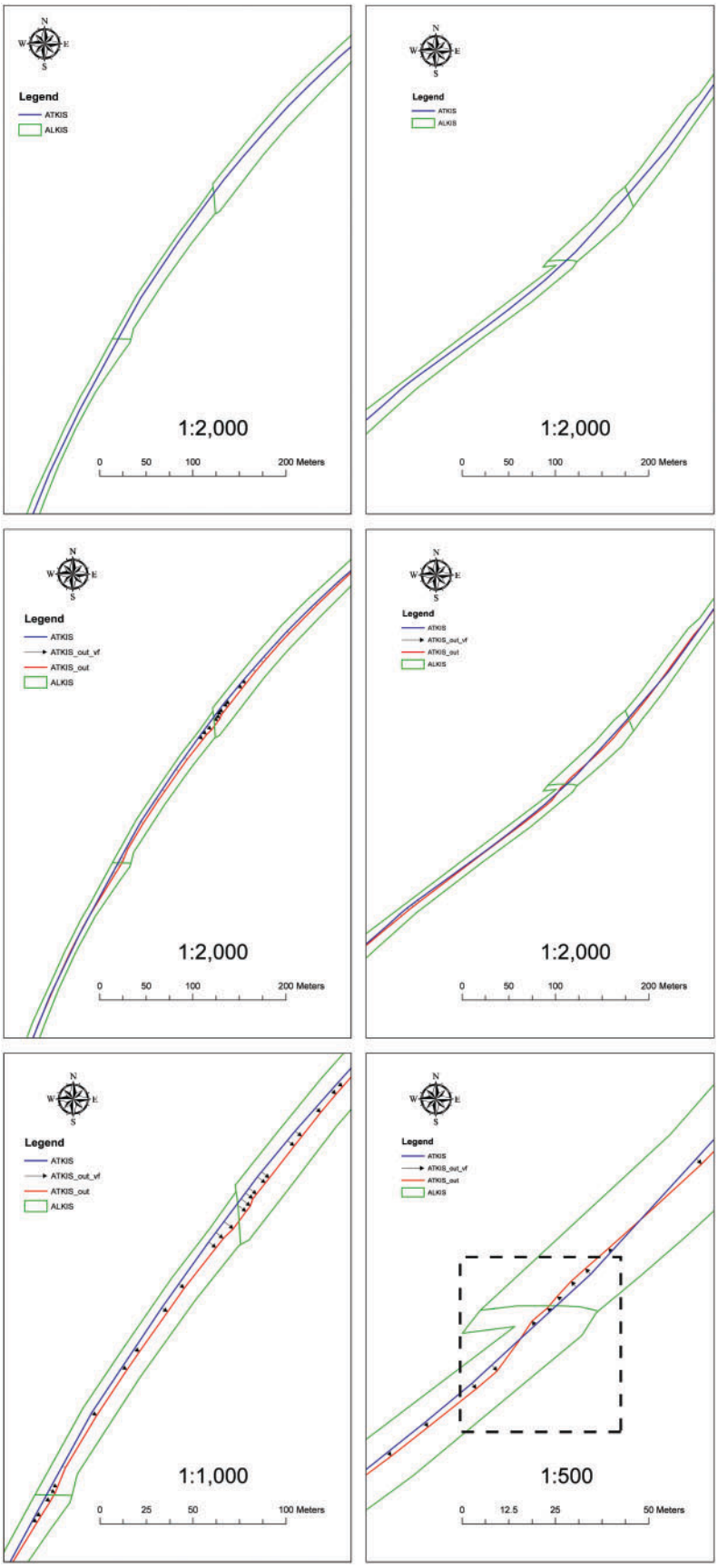
It should be noted that though several source polyline vertices exist outside the reference polygon (boxed on left example, top), these are not shifted toward the polygon due to the use of the direction (azimuth) constraints discussed in the previous section. On the right example, the upper-left branches are composed of two separate features; this leads to an ambiguity in the correspondences for the source vertices in that area. As can be seen by the results, the final polyline is aligned only to the lower geometry (feature), where in this case the desired output might suggest otherwise. It is possible to implement a pre-processing stage using a dissolve function on all separate features that will solve such ambiguity.

Figure 9 depicts two extracts of the output after implementing the proposed methodology on the dataset presented in Fig. 1 (rectangular areas): the top row is the initial state superimposed with the resulting vector fields, where the bottom row depicts the final positioning. Both examples show precise alignment results of the final geometry, e.g. polyline network that is shifted inside the area bounded by the reference polygon, even in areas where the source polyline exists outside the bounding areas of the reference polygons. In both cases the proposed methodology was able to monitor these irregularities in the geodata, while aligning both geometries precisely.

It is worth noting that on a global scale the results do not change or deviate when the number of features to be aligned is modified, i.e. data amount is increased or data coverage is expanded. The geometric adjustment results do not differ since the strategy presented here exploits only local topologic and geometric relationships that exist between corresponding features - prior to the implementation of LSA, i.e. no global transformation and alignment is performed during the process. Minor geometrical alterations might exist, but these are only local ones that have a restricted affect on the overall solution.

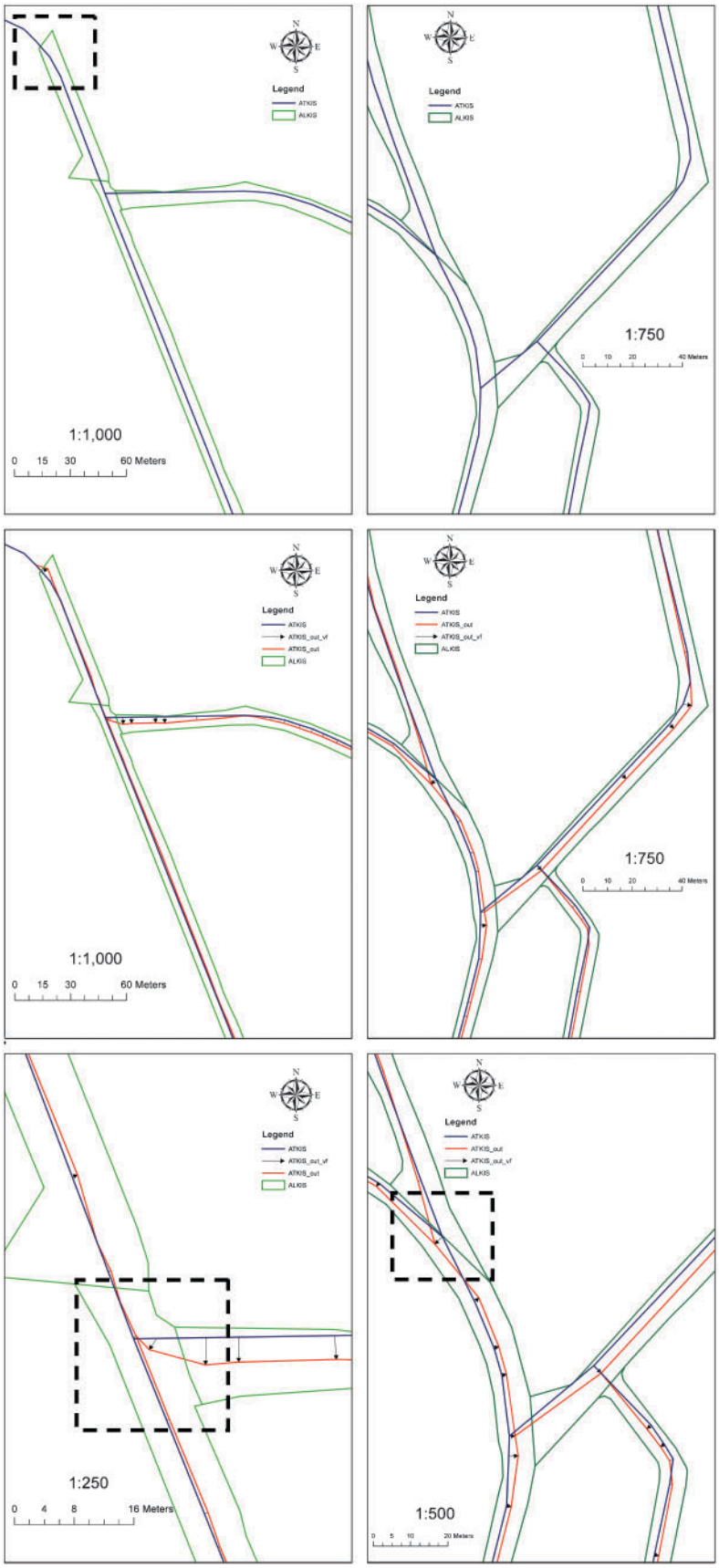
Comparison of results with straight skeleton formation

In order to evaluate the outcome, a comparison was made with the straight skeleton process as presented in



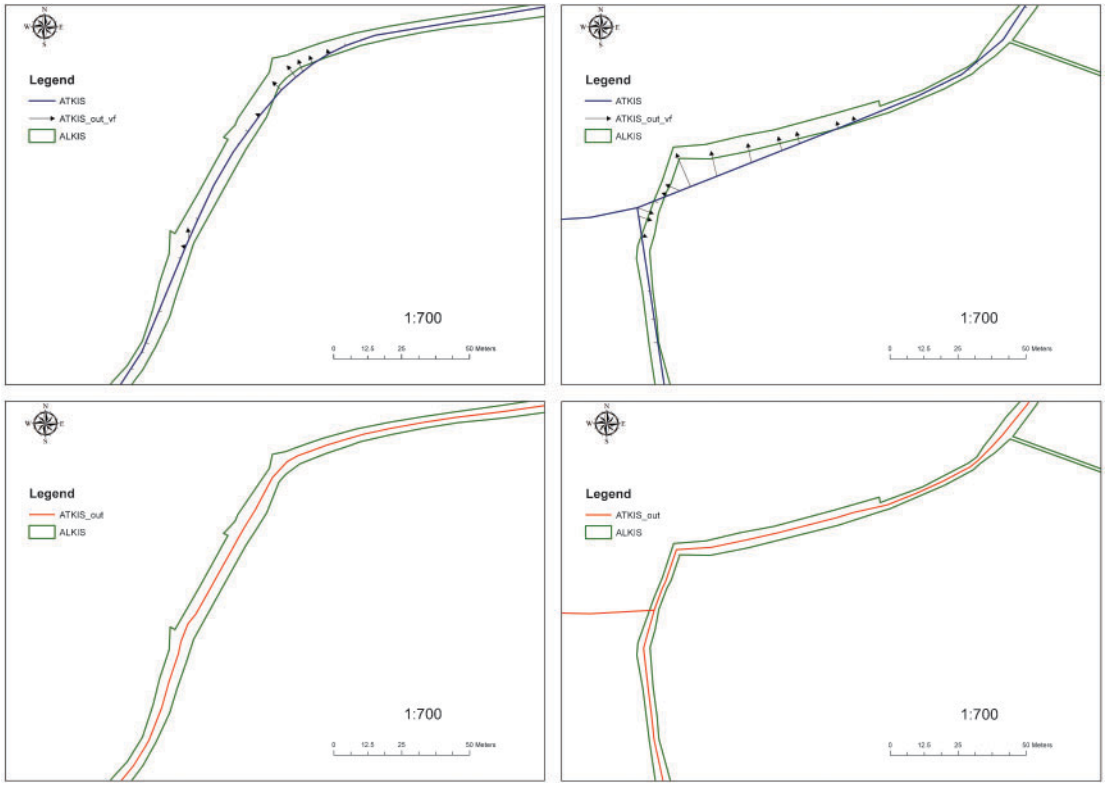
7 Example of two processes (left and right): initial state (top), final state (middle), and zoomed area (bottom), showing full geometrical alignment

[**11**]. The results are depicted in Fig. 10. It is visible that in areas defined by clear boundary of (close-to-) parallel polygon edges, both algorithms produce the same results, with slight displacement positioning of several centimetres. Still, in other areas where polygon borders are not clearly defined, as in junctions and abrupt



8 Example of two processes (left and right): initial state (top), final state (middle), and zoomed area (bottom), showing full geometrical alignment in vicinity of junction

changes (such as parking areas in polygon features), the straight skeleton algorithm does not produce results or produces ambiguous ones – sometime even producing data-outliers (bottom example). In contrast, our proposed algorithm manages to overcome the ambiguous geometry, producing qualitative results that satisfy the



9 Example of two area extracts from Fig. 1 (left and right): initial state (top), and final state (bottom), showing source polyline features that are 'inserted' into bounding area of the reference polygons with full geometrical alignment

aim of aligning these different networks of different structural geometric primitives.

Closing discussion

Previous paragraphs presented qualitative analyses of the results with quantitative ones. It is important to note that precise statistical analysis is hard to obtain in this case, since no 'ground truth' or 'proper' reference is given for such an evaluation (this also relates to the comparison presented in the previous paragraph). As such, statistical tests and analysis alone will not suffice and work here, such that qualitative results are the main means on which to rely.

Conclusions and future work

A geometrical adjustment approach for the alignment and harmonization of vector databases was presented. The aim here is to handle geodata from databases where each is represented by different geometric primitives; the emphasis and case study presented here was given to road networks existing in geodatabases. This strategy exploits local topologic and geometric relationships between corresponding line features prior to the implementation of LSA, and observes local distortions and ambiguities that might exist; as opposed to a global transformation and alignment. As depicted and analyzed in the examples, the outcome presents a significant improvement on the initial geometry, while specific constraints are tailored with respect to the different structures. The statistical analysis and comparison to an alternative algorithm proved that the approach presented is statistically sound, presenting a qualitative and reliable solution to the problem of spatial inconsistency that typically occurs when different vector geodatabases are overlaid.

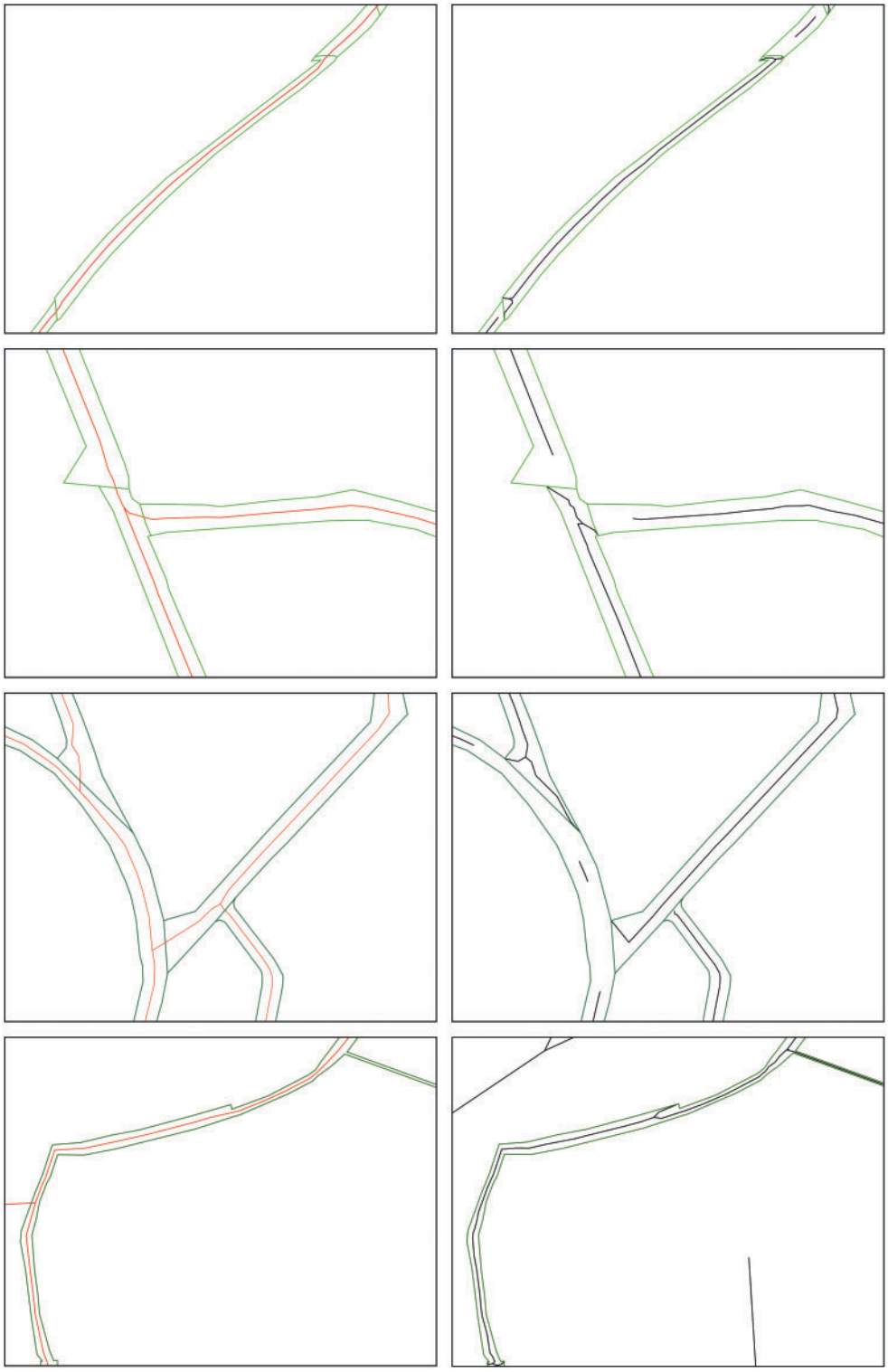
Future work will entail adding more constraints to fine tune problems that are encountered from poor geometrical data constellation. Also, additional geometrical constrains might be considered to be integrated into the LSA model that aim at preserving specific geometrical attributes in the output (for example: keeping straightness of source data or preserving

Table 1 Statistical values of railway path alignment

	First iteration (v^tv)	Second iteration (v^tv)
ATKIS 202 out of 325 points	713·10 m^2	0·20 m^2
ALKIS 105 out of 228 points	0·01 m^2	0·00 m^2

Table 2 Statistical values of intersection railway path alignment

	Left area (v^tv)	Right area (v^tv)
ATKIS	164 out of 190 points 803·60 m^2	46 out of 151 points 142·60 m^2
ALKIS	90 out of 323 points 0·00 m^2	83 out of 191 0·10 m^2



10 Comparing outcome of the proposed algorithm (left) – red polyline after alignment, and straight skeleton algorithm (right) – black line representing skeleton geometry (superfluous line features created are outliers)

directional constraints in the network). Larger databases that have wider coverage areas will also be analyzed, which will mainly have an effect on the size of the normal equation system to be solved. In order to reduce the number of features involved, a possibility could be to apply a hierarchical partitioning of the space, e.g. using the major road network as objects on the high level, and subsequently adjust the features within such a high level network mesh.

There are also cases, where a mere geometric analysis of possible corresponding features is not successful, as several neighbouring objects are possible. Therefore, in the next stage we will make use of supplementary data to 'enrich' the current geometric process, e.g. exploit existing semantics or calculate geometric attributes and other topological characteristics, to better identify corresponding features for alignment.

The general purpose of the application we propose is to adapt elongated objects in different datasets with multiple representations. In a first step the process could be applied to the update process of spatial data

infrastructure. With the method it becomes possible to sketch the centreline of a road and then to adjust it to a high quality road polygon. It is also possible to add restrictions (constraints) to the line, such as that the bending should not be larger than a certain threshold. Such restrictions are useful to obtain a correctly generalised representation of the object. In a later step the method could be used to adapt not only geometries of multiple represented objects but also to transfer semantic information.

Vice versa, the discrepancies can be used as a similarity measure to compare linear and polygonal objects. Based on this measure the similarity of the semantic of the objects could be specified. However, further research should be done to prove the ability of such an approach.

References

1. Aho, A. *et al*., 1987. Data Structures and Algorithms. Addison-Wesley Series in Computer Science and Information Processing, 427 .
2. Boljen, J., 2010. Geometrischer Abgleich von Vektordaten auf der Basis eines Gauss-Helmert-Modells, *ZfV Zeitschrift für Vermessungswesen*, 137: 314–323.
3. Butenuth, M., Gösseln, G. V., Tiedge, M., Heipke, C., Lipeck, U., Sester, M., 2007. Integration of Heterogeneous Geospatial Data in a Federated Database. *ISPRS Journal of Photogrammetry and Remote Sensing*, 62(5): 328–346.
4. Cobb M. A., Chung M. J., Foley H., Petry, F. E., Shaw K. B., Miller H. V., 1997. A Rule-Based Approach for the Conflation of Attributed Vector Data. *Geoinformatica*, 2: 7–36.
5. Dalyot, S., Dahinden, T., Schulze, J. M., Boljen, J., Sester, M., 2012. Geometrical Adjustment Towards the Alignment of Vector Databases, ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences, I–4, 13–18.
6. Doytsher, Y., Filin, S., Ezra, E., 2001. Transformation of Datasets in a Linear-based Map Conflation Framework, *Surveying and Land Information Systems*, 61(3): 159–169.
7. Filin, S. and Doytsher, Y., 2000. A Linear Conflation Approach for the Integration of Photogrammetric Information and GIS Data, International Achives of Photogrammetry and Remote Sensing, XXXIII, Part B3.
8. Gabay, Y. and Doytsher, Y., 1994. Automatic adjustment of line maps, *Proceedings of the GIS/LIS'94 Annual Convention*, Arizona, Phoenix, USA: 333–341.
9. Göpfert, J. and Rottensteiner, F., 2010. Using Building and Bridge Information for Adapting Roads to ALS Data by Means of Network Snakes, In: Int. Arch. PhRS, Vol. XXXVIII, part 3A, Paris, 163–168.
10. Göpfert, J. and Rottensteiner, F., 2012. Network Snakes for Adapting GIS Roads to Height Data of Different Data Sources – Performance Analysis using ALS Data and Stereo Images. ISPRS Annals of the Photogrammetry, *Remote Sensing and Spatial Information Science*, I-3, 209–214.
11. Haunert, J.-H. and Sester, M., 2008. Area Collapse and Road Centerlines based on Straight Skeletons, *GeoInformatica*, 12(2): 169–191.
12. Ogniewicz, R. and Ilg, M., 1992. Voronoi skeletons: theory and applications. *Computer Vision and Pattern Recognition*, 1992. Proceedings CVPR '92., 1992 IEEE Computer Society Conference, 63–69.
13. Oswin, A. *et al*., 1995. A Novel Type of Skeleton for Polygons. *Journal of Universal Computer Science*, 1(12): 752–761.
14. Rosen, B. and Saalfeld, A., 1985. Match Criteria for Automatic Alignment, Proceedings of Auto-Carto VII, 1–20.
15. Siriba, D. N., 2011. Conflation of Provisional Cadastral and Topographical Datasets. Ph.D. Thesis. Zugl. Hannover, Univ., Diss., 2011. (DE-101)1021254444.
16. Vassilaki, D. *et al*., 2009. Multitemporal data registration through global matching of networks of free-form curves. *FIG Working Week 2009*, Eilat, Israel, 3rd to 8th May. 18.
17. Vassilaki, D. *et al*., 2012. Automatic ICP-Based global matching of free-form linear features. *The Photogrammetric Record*, 27(139): 311–329.
18. Volz, S., 2006. An iterative approach for matching multiple representations of street data. *International Archives of Photogrammetry*, Remote Sensing and Spatial Information Sciences 36 (Part 2/W40), 101–110.
19. Walter, V., 1997, Zuordnung von raumbezogenen Daten – am Beispiel der Datenmodelle ATKIS und GDF. *Ph.D. thesis,* Deutsche Geodätische Kommission (DGK), Reihe C, Heft Nr. 480, 127.
20. Walter, V. and Fritsch, D., 1999, Matching Spatial Data Sets: a Statistical Approach, *International Journal of Geographical Information Science*, 13(5): 445–473.
21. Xiong D., 2000. A three-stage computational approach to network matching, Transportation research part C, 71–88.