Development of a data structure and tools for the integration of heterogeneous geospatial data sets

Butenuth, M. (1); Gösseln, G. v. (2), Heipke, C. (1), Lipeck, U. (3); Sester, M. (2); Tiedge, M. (3)
(1) Institute of Photogrammetry and GeoInformation, University of Hannover,
Nienburger Str. 1, 30167 Hannover, Germany,
{butenuth, heipke}@ipi.uni-hannover.de
(2) Institute of Cartography and Geoinformatics, University of Hannover,
Appelstr. 9a, 30167 Hannover, Germany
{goesseln, sester}@ikg.uni-hannover.de
(3) Institute of Practical Informatics, University of Hannover,
Welfengarten 1, 30167 Hannover
{ul, mti}@dbs.uni-hannover.de

Abstract

The integration of heterogeneous geospatial data sets offers extended possibilities of deriving new information which could not be accessed by using only single sources. Different acquisition methods, data schemata and updating periods of the topographic content leads to discrepancies in geometry, accuracy and topicality which hampers the combined usage of these data sets. The integration of different data sets – in our case topographic data, geoscientific data and imagery – allows for a consistent representation, the propagation of updates from one data set to the other and the automatic derivation of new information. In order to achieve these goals, basic methods for the integration and harmonisation of data from different sources and of different types are needed. To provide an integrated access to the heterogeneous data sets a federated spatial database is developed. We demonstrate two generic integration cases, namely the integration of two heterogeneous vector data sets, and the integration of raster and vector data.

1. Introduction

Geospatial data integration is often applied to solve complex geoscientific questions. To ensure successful data integration, i.e. ensure that the integrated data sets fit to each other and can be analysed in a meaningful way, an intelligent strategy is required due to the fact that these data sets are mostly acquired using different methods, quality standards and at different points in time. Differences between printed analogue maps were not as apparent as are those of digital data of today, when different data sets are overlaid in modern GIS-applications. Integrating different data sets allows for a consistent representation and for the propagation of updates from one data set to the other.

To enable the integration of vector data sets, a strategy based on semantic and geometric matching, object based linking, geometric alignment, change detection, and updating will be used. With this described strategy the actual topographic content from an up-to-date data set can be used as a reference to enhance the content of certain geoscientific data sets. In addition, the integration of two data sets with the aim to derive an updated data set with an intermediate geometry based on given weights is possible. The integration of raster and vector data sets is the second integration task dealt with in this paper. As an example, field boundaries and wind erosion obstacles are extracted from aerial imagery exploiting prior GIS knowledge. One application area are geoscientific questions, for example the derivation of potential wind erosion risk fields, which can be generated with field boundaries and additional input information about the prevailing wind direction and soil parameters. Another area is the agricultural sector, where information about field geometry is important for tasks such as precision farming or the monitoring and control of subsidies.

The paper is structured as follows: The following section gives an overview of the state of the art concerning the topic of data integration. Afterwards, the used data sets are presented and an architecture for database supported integration is described. Methods for the integration of vector/vector and raster/vector data integration are highlighted in the following section. Results demonstrate the potential of the proposed solution, finally a set of conclusions is given and further work is discussed.

2. State of the art of geospatial data integration

The integration of vector data sets presented in this paper is based on the idea of comparing two data sets, while one is used as a reference and a second one – the candidate – is aligned to the first one, which is a general matching problem, see e.g. Walter and Fritsch (1999). For the integration of multiple data sets, it has been shown how corresponding objects can be found when several data sets have to be integrated (Beeri et al., 2005). Due to the complexity of the integration problem it is very difficult to solve this task with one closed system, therefore the development of a strategy based on component ware technology was proposed (Yuan and Tao, 1999) and a software prototype for the vector data integration has been developed as a set of components to ensure the applicability in different integration tasks. While this approach uses a reference data set to enhance and update the topographic content of a candidate data set, data integration can also be used for data registration, when one data set is spatially referenced and the other has to be aligned to it (Sester et al., 1998). In order to geometrically adapt data sets of different origin, rubber sheeting mechanisms are being applied (Doythser, 2000). Strategies applied to cadastral data based on triangulation to enhance the rubber-sheeting process have been presented by Hettwer and Benning (2000).

The recognition of objects with the help of image analysis methods starts often with an integration of raster and vector data, i.e. using prior knowledge to support object extraction. An integrated modelling of the objects of interest and the surrounding scene exploiting the context relations between different objects leads to an overall and holistic description (Baltsavias, 2004). In this paper, the extraction of field boundaries and wind erosion obstacles from imagery is chosen to demonstrate the methodology integrating raster and vector data. In the past, several investigations regarding the automatic extraction of man-made objects have been carried out (e.g. Mayer, 2001). Similarly, the extraction of trees has been accomplished, cf. Hill and Leckie (1999) for an overview of approaches suitable for woodland. In contrary, the extraction of field boundaries is not in an advanced phase: a first approach to update and refine topologically correct field boundaries by fusing raster-images and vector-map data is represented in Löcherbach (1998). The author focuses on the reconstruction of the geometry and features of the land-use units, however, the acquisition of new boundaries is not discussed. In Torre and Radeva (2000) a so called region competition approach is described, which extracts field boundaries from aerial images with a combination of region growing techniques and snakes. To initialise the process, seed regions have to be defined manually, which is a time and cost-intensive procedure.

In order to connect heterogeneous databases, first so-called multi-database architectures had been discussed for loose coupling. Subsequently, so-called federated databases have been chosen to support closer coupling (Conrad, 1997). Federated databases allow integrating heterogeneous databases via a global schema and provide a unified database interface for global applications. Local applications remain unchanged, as they still access the databases via local schemata. For database schema integration a broad spectrum of methods has been investigated (Batini et al., 1986), but identifying objects is typically restricted to one-to-one-relationships. In context of geospatial integration more sophisticated methods are needed, to incorporate complex correspondences between objects (many-to-many-relationships), which usually are not considered in federated databases. Whereas there are a lot of overview articles of spatial databases (e.g. Rigeaux, 2002), federated spatial databases are hardly investigated with the exception of (Devogele, 1998; Laurini, 1998).

3. Architecture for integration

Different geospatial data sets which represent the same real world region, but cover different thematic aspects, are acquired with respect to different needs. In this section we present an architecture that provides an integrated access to heterogeneous data sets. It is designed to store and export results of the vector/vector and the raster/vector integration steps. This task is accomplished according to the paradigm of federated databases. For this purpose the known architecture of a federated database is expanded to handle geospatial data. In order to select certain objects satisfying given semantic criteria it is possible to define mappings to harmonise the attributes of the different data sets. Furthermore, the database provides mechanisms to pre-process geospatial objects for the integration of raster and vector data. Fig. 1 gives a simplified overview of the realised system architecture with respect to the interaction between the federated database and the integration process, namely object matching and extraction.

In the next section, the involved vector and raster data sets are described to demonstrate how much the geospatial data models differ structurally and semantically. Then the architecture and modelling concepts of the database integration are explained; they provide an organisational framework for the approaches of geospatial data integration given in section 4.

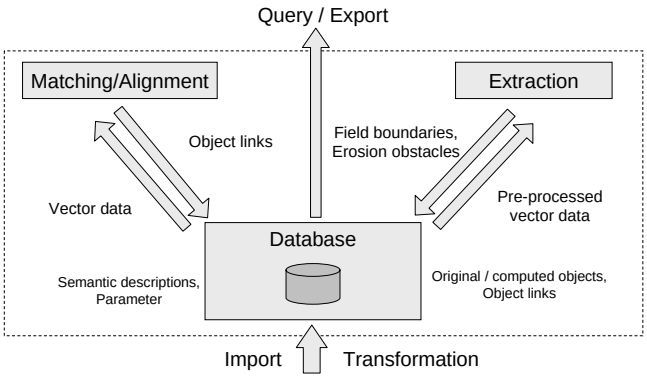


Fig. 1. System overview.

3.1 Used Data sets

The vector data sets used in this project include the German topographic data set (ATKIS DLMBasis), the geological map (GK) and the soil science map (BK), all at a scale of 1:25000. Simple superimposition of different data sets already reveals some differences. These differences can be explained by looking at the creation the maps. For ATKIS the topography is the main thematic focus, for the geoscientific maps it is either geology or soil science. Thus, these maps have been produced using the result of geological drilling, and according to this punctual information, area objects have been derived using interpolation methods based on geoscientific models. They are, however, related to the underlying topography. The connection between the data sets has been achieved by using the topographic information together with the geoscientific results at the point of time, when the geological or soil science information was collected. The selection and integration of objects from one data set to another one was performed manually and in most of the cases the objects have been generalised by the geoscientist. While the geological content of these data sets keeps its topicality for decades, the topographic information in these maps does not: In general, topographic updates are not integrated unless new geological information has to be inserted in these data sets. The geoscientific maps have been digitised to use the benefits of digital data sets, but due to the digitalisation even more discrepancies occurred. Another problem which amplifies the deviations of the geometry is the case of different data models. Geological and soil science maps are single-layered data sets which consist only of polygons with attribute tables for the representation of thematic and topographic content, while ATKIS has a multi-layered data structure with objects of all geometric types, namely points, lines and polygons, equally with attribute tables. In addition to the described vector data, raster data sets are used to enable object recognition while exploiting the prior ATKIS knowledge. The raster data sets are aerial images or high resolution satellite images, which include an infrared channel.

3.2 Architecture and concepts of integration

As the previous section has shown, the various geospatial data sets differ significantly due to the various objectives of their acquisition. In order to integrate the corresponding databases we have chosen the architectural paradigm of federation (Conrad, 1997), as it gives a close coupling at the same time and keeps the databases autonomous. Hereby, the matching and extraction processes are given an integrated view to the different databases via a global database schema (global applications). Nevertheless, particular applications (like import and export processes) may still access the databases locally as shown in Fig. 2.

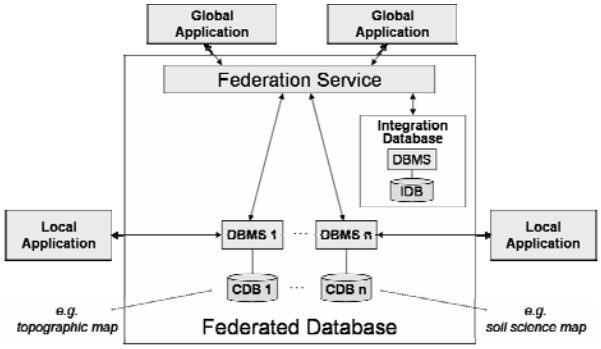


Fig. 2. Architecture of a federated database.

The federation service requires an “integration database” (cf. section 3.2.4) on its own to maintain imports and descriptions of the involved data sets (component databases), and to incorporate qualified links between object as the result of the matching process as well as further findings such as geometric adjusted and new extracted objects.

3.2.1 Schema adaptation

To make the structurally different data sets accessible to the federation service a generic but flexible export schema was designed based on experiences with geospatial data sets containing topographic objects with respect to object-relational databases (Kleiner, 2000). The schema contains all objects, object classes, attribute types and attribute values, each of them in one entity type (or table in the relational DBMS). Fig. 3 shows the schema for topographic data (ATKIS), the geoscientific data sets get isomorphic export views; in more detail they have application-specific attribute types and object classes according to their own representation model.

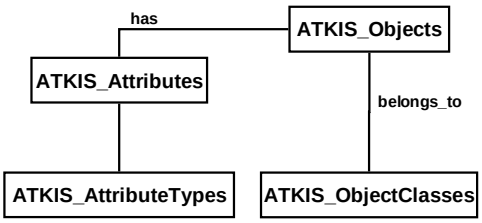


Fig. 3. Export schema for the topographic map (ATKIS).

A geoobject of entity type ATKIS_Objects, e.g. a road, has several entries of type ATKIS_Attributes, namely (attribute, value)-pairs like e.g. (width, 10 meters). The corresponding type of the attributes or the classification of the geoobjects can be found in the collections ATKIS_AttributeTypes and ATKIS_ObjectClasses.

3.2.2 Object linking

Given the structural adaptation of the different data sets, the federated database can be enabled to incorporate correspondences through so called links. Linking objects, however, should not only involves simple one-to-one-relationships, as real-world objects are represented differently with respect to different maps. The federation service has to cope with more complex correspondences namely one-to-many- and even many-to-many-relationships as shown in Fig. 4, which represents different partitions of a real world object in two maps. This task is accomplished with a flexible schema, that integrates these general correspondences as attributed one-to-one links between aggregated objects. Fig. 4 shows an instance of three and two objects, respectively, e.g. a section of a water body segmented in two different ways, whose aggregations (denoted by dashed lines) are linked.

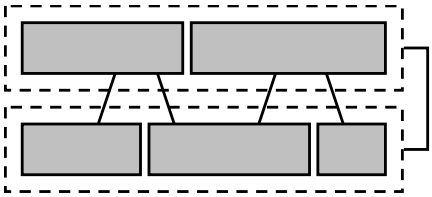


Fig. 4. Realisation of a many-to-many-relationship as a link between object aggregations.

3.2.3 Attribute harmonisation and semantic selection

In order to provide the applications with a model independent and uniform method to access certain objects with respect to thematic attributes, a mechanism for the semantic description of geoobjects was developed, to characterise comparable object sets for the matching process and to characterise object selection for the extraction process. To fulfil these requirements, the architecture of federated databases had to be expanded to unify the handling of semantic descriptions. Fig. 5 shows two simplified semantic selections of topographic objects, namely of open landscape and a partitioning network.

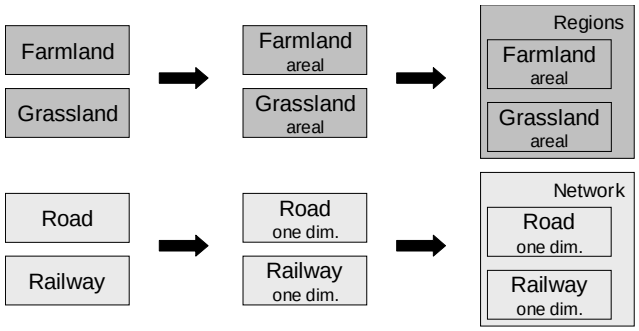


Fig. 5. Semantic selections for regions and networks.

Semantic object selections are defined in the following three stages: Coarse semantic classification is achieved through the references to object classes given by the export views. Fig. 5 depicts some object classes of the topographic map, e.g. farmland and roads. Next, a more precise characterisation is provided through the specification of object attributes, i.e. the coarse selection via object classes is restricted by attribute conditions. For instance, road objects appear as both one-dimensional and two-dimensional objects due to acquisition rules. In order to build a partitioning network only the one-dimensional road objects are needed. Finally, fine object classes are merged to class sets, which provide semantic selections for the global applications, independent of the original data set's semantic specifications. Next to the structural unification through export views attribute harmonisation is achieved by connecting two conforming semantic selections of two different data sets (e.g. water bodies both in the topographic and the geological map). It is necessary to provide this semantic description for any representation model only once, independent of the quantity of instances of this particular model (component databases).

3.2.4 Integrated schema

Fig. 6 summarises the schema architecture of the integration database. The component databases are both original involved geospatial data sets based on the previously described export views, and the term "Objects" stands for all objects of the integration database, i.e. adjusted and extracted geometric objects. The different parts of Fig. 6 show that the federation service is supported with respect to the following tasks for

- the *model description*, characterisation of object classes and attribute types of a certain data set model
- the *registration*, registering the component databases
- the *semantic selection* as described in the previous section
- the *application control*, which stores meta data about extraction and matching processes, in particular about the used semantic selections, and links between the involved component databases
- the *linking* objects from different data sets (object linking, cf. Section 3.2.2)

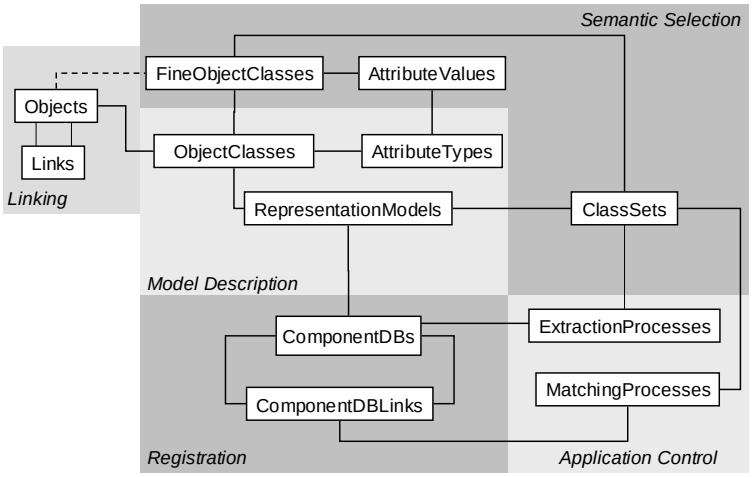


Fig. 6. Overview of the integrated schema.

4. Methods of data integration

In this section the methodologies of the vector/vector and the raster/vector data integration are described. First, the integration of heterogeneous vector data sets which have been acquired for different purposes and with unequal updating strategies is presented based on a component based strategy. Subsequently, the integration of raster and vector data is highlighted with the example of the extraction of field boundaries and wind erosion obstacles from imagery exploiting prior GIS knowledge.

4.1 Integration of vector data

At the beginning of the integration process the semantic content of all data sets was compared. According to this step, certain *selection groups* were built up for each data set (e.g. water area). This selection is mandatory to avoid comparing "apples and oranges" and has to be the first step to ensure a successful integration. An area-based matching process is used for the creation of links between object candidates. These links are stored in the federated database using a XML-schema, followed by an alignment process which reduces geometric discrepancies to a minimum to ensure satisfying results in the subsequent intersection process, but will still be capable of deciding between geometric discrepancies based on map creation or topographic changes which occurred during the different times of acquisition. A rule-based evaluation of the intersection results is used for change detection.

4.1.1 Revelation of links between corresponding objects

Various data sets have different forms of representations for certain topographic objects (e.g. rivers), the decision which kind of representation to take often depends on specific attributes, e.g. in (ATKIS DLMBasis, cf. Section 3) the width of the river is used for this decision, thinner than 12 meters – polyline, wider than 12 meters polygon. Due to the fact that there are different thresholds for each data set, these differences have to be resolved using harmonisation strategies. To ensure a suitable result in the revelation of links, line objects have to be transformed into polygons by applying a buffer algorithm using the width attribute.
Another problem is the representation of *grouped objects* in different maps. For a group of water objects, e.g. a group of ponds, the representation in the different data sets could either be a group of objects with the same or a different number of objects, or even a single generalised object (see Fig. 7). Finally, also objects can be present in one data set and not represented in the other. All these considerations lead to the following relation cardinalities that have to be integrated: 1:0, 1:1, 1:n, and n:m. After the corresponding relations have been identified, each selection set will be aggregated, so they can be handled as 1:1 relations, so called *relation-sets* (Goesseln and Sester, 2004).

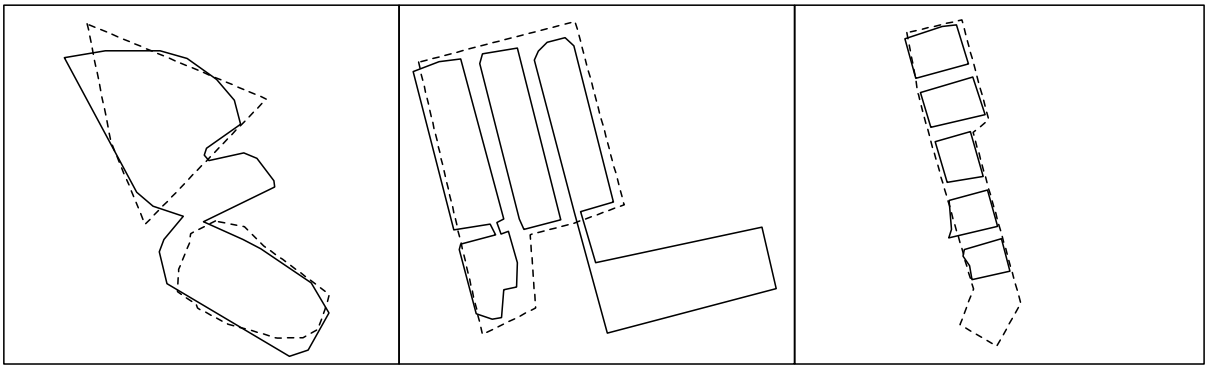


Fig. 7. Different representations - ATKIS (solid line), GK (dotted line).

These *relation-sets* will be visualised to the operator - using a GUI based application - enabling a manual correction of the derived links. With this software each *relation-set* can be inspected and edited, to check whether the automated process has failed to build up the suitable correspondences between the selected data sets. Because of to the fact that the objects from all three data sets are representations of the same real world objects, they show apparent resemblance in shape and position. Nevertheless the alignment of the geometries is required after the evaluation of the matching results. As it will be described later, there are different geometric alignment method required for covering all alignment tasks. Therefore the technique offering the most suitable result can be selected for every single *relation-set*.

4.1.2 Geometric Alignment of corresponding objects

Objects which have been considered as a *matching pair* could be investigated for change detection using intersection. At this stage the mentioned differences will produce more problems which are visible as discrepancies in position, scale and shape. These discrepancies will lead to unsatisfying results in the evaluation of the resulting elements almost and this would evoke an immoderate estimation of the area investigated as change of topographic content. Therefore a geometric adaptation will be applied, leading to a better geometric correspondence of the objects. For these adaptation processes thresholds are required which allow the reduction of discrepancies which are based on map creation, but will not cover the changes which happened to real world objects between the different times of data acquisition.

Iterative closest point (ICP)

The iterative closest point algorithm (ICP) developed by (Besl and McKay, 1992) has been implemented to achieve the best fitting between the objects from ATKIS and the geo-scientific elements using a rigid 7 parameter transformation. The selection of a suitable algorithm used for ICP is depending on the alignment to be performed, in this case the problem is reduced to a 2D problem requiring four parameters (position, scale and orientation) an solved using a Helmert-transformation. These calculations are repeated iteratively and will be evaluated after each calculation; the iteration stops when no more variation in the four parameters occur. At the end of the process the best fit between the objects using the given transformation is achieved. Evaluating the transformation parameters allows for classifying and characterising the quality of the matching: in the ideal case, the scale parameter should be close to 1 and rotation and translation should be close to 0. Assuming, that the registration of the data sets is good, these four parameters exactly meet the reasons for the integration of analogue produced data sets, that have been created by manual copying of printed maps. Therefore a greater scale factor can be an indicator for differences between two objects that are not based on map creation, but on a change on the real world object, that occurred between the different times of data acquisition (Goesseln and Sester, 2004). At the end of the process the best fit between the objects using the given transformation is achieved. The result of this transformation is stored as a set of shifting vectors, which are required in a subsequent step in which the neighbourhood of the transformed objects will be aligned. This step will be described later on (cf. Section 4.1.3). The application of the iterative adaptation using the ICP approach based on Helmert-transformation showed very good results and revealed the possibility of reducing the amount of objects which have to be evaluated manually. However there are some situations where this approach does not generate sufficient results (e.g. objects which cover several map-sheets or at least touch the map boundaries).

Dual interval alignment (DIA)

The DIA approach has been implemented, enabling the alignment of local discrepancies of corresponding geometries by calculating the transformation of single vertices, based on the ideas of Kohonen (1997), however this approach handles each vector separately. Corresponding objects which have been assigned as representations of the same real world object through the matching process are investigated based on their vertices. For every point in one object the nearest neighbour in the corresponding partner object is determined using the criterion of proximity. The conformation approach evaluates the distance between these coordinates, based on an interval which is predetermined by the human operator. This threshold defines the largest distance – representing a change in geometry – which will be suitable for the candidate data set. Distances exceeding this threshold implicate a topographic difference which has to be investigated during field-work.

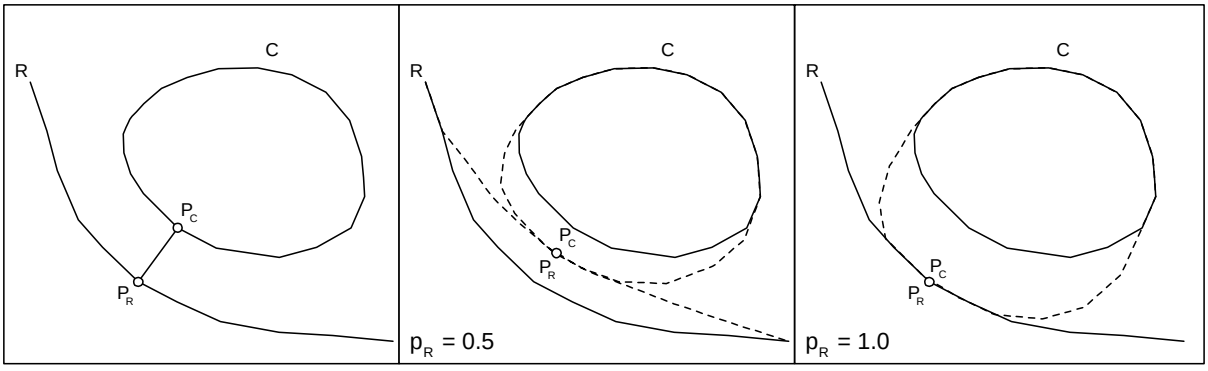


Fig. 8. Application of DIA for the partial alignment of object geometries (schematic).

As it can be seen in Fig. 8, for each point (P_C) from object C and the corresponding point (P_R) of the linked object R, the point transformation is calculated based on the euclidean distance (d) between these points. The new coordinates are determined taking interval ranges a and b into account. Points within the first distance

interval (*0<d<a*) are aligned to a single point, a distance falling into the second interval (*a<d<b*) will lead to an approximation of the selected points. Points with a distance beyond b will not be adapted (see Eq. 1).
It seems to be a paradox that the complete alignment must not be the perfect result. The integration of data sets which cover the same area, which are based on the same method of representation and are acquired at nearly the same point of time, can be performed by using an alignment strategy with elimination of all differences. While integrating data sets which have been acquired at different points in time it is obvious that a certain amount of change to topography, built-up area and/or vegetation has occurred. Therefore an alignment threshold is required which allows the operator to decide between errors due to map creation or real-world changes. The introduction of a second threshold follows the idea of "fuzzy" logic and ensures that there are no points of hard discontinuities at the geometry of aligned objects.
The integration of a weight *p* (see Eq. 1) to the alignment process does not only take different accuracies of geometries into account, but opens this approach to a much wider range of conflation tasks. E.g. in a project where one data set is handled as a reference data set and must not be changed (weight set to 1). In other cases, when two data sets have to be aligned and no single data set can be selected as reference the alignment is performed using the common idea of conflation by aligning two data sets to a new improved data set.

$$(x,y)_{P_{new}} = \begin{cases} (x,y)_{P_{old}} + (\pm\Delta(x,y)\cdot p) & ,0 < d < a. \\ (x,y)_{P_{old}} + (\pm\Delta(x,y)\cdot p)\cdot f(d) & ,a < d < b. \\ (x,y)_{P_{old}} & ,b < d. \end{cases} , \qquad (1)$$

$$f(d) = \left(\frac{b-d}{b-a}\right), \quad p(P_c) = 1 - p(P_r), \quad \Delta_{RC}(x,y) = (x,y)_R - (x,y)_C$$

Calculating the shift distance based on the nearest neighbour enables very good alignment, but can result in topologic errors. This requires the integration of an additional criteria which is vertex orientation. Therefore the orientation of the polygon segments will be calculated for all corresponding objects. If a point and its corresponding partner are selected using the distance criterion, the direction to each corresponding successor will be calculated. If the difference between these directions exceeds a given threshold, the points must not be aligned due to the assumption that they do not represent the same "side" of the real world.
Comparing the adaptation approaches ICP and DIA, each is suitable for a different kind of objects in this project. ICP matches the idea, that the majority of the geometric discrepancies is caused in the way the data sets have been created by integrating topographic elements through manual copying. The resulting parameters can be used for the investigation and the evaluation of the influences which were responsible for the geometric discrepancies: An object which can be aligned by just using translations with a small change in scale can be judged as minor error based on manual copying. A larger scale factor can reveal topographic changes on the real world objects, which have to be investigated by a human operator.
The transformation, the implemented ICP algorithm is based on, is very fast and reliable, but depending on the chosen transformation algorithm it does not give satisfying results for larger, irregular shaped objects like rivers, or objects that have been changed during different periods of time and therefore only match partially. But as good as the alignment results of DIA are, it is much more time consuming and susceptible to errors. The combination of both approaches delivered very good results, offering the possibility to assess the geometric discrepancies by evaluating the resulting ICP-parameters, and aligning large object groups or partially matching objects using DIA. The automatic guided decision between these methods has not been completed yet. So far both methods will be applied for every *relation-set* and the most suitable result will be chosen by comparing the results with certain geometric operators (e.g. angle histogram, symmetric difference).

4.1.3 Neighbourhood adaptation using rubber-sheeting

The individual alignment of selected objects would result in gaps, overlaps or inconsistencies concerning the rest of the data set, so that the neighbourhood of the aligned objects must be transformed equivalent. To ensure an overall alignment, the results which originate from the individual alignment processes are stored as a collection of displacement vectors. All vectors will build up a vector field which is the basis of the neighbourhood adaptation ensuring a homogeneous data set. Using a distance weightened interpolation the rubber-sheeting method calculates a new transformation target for every point in the data set based on the vectors derived from the alignment.
This strategy has to be carefully adapted for every adaptation process regarding to the used data set. Different data sets require different constraints the rubber-sheeting algorithm must be able to take into account. These constraints can be e.g. points or areas which must not be changed, like fixed points or areas which have been updated manually in advance, or objects of higher category.

4.2 Integration of raster and vector data

The integration of raster and vector data is highlighted by means of the extraction of field boundaries and wind erosion obstacles from imagery exploiting prior GIS knowledge. First, the integrated modelling and the derived strategy are described, followed by the presentation of fully automatic methods to extract the field boundaries and wind erosion obstacles.

4.2.1 Model and strategy

The semantic model comprises the integration of raster data (imagery) and vector data (GIS data) as starting point for the object extraction, as described in detail in Butenuth (2004). The model is differentiated in an object layer, a geometric and material part, as well as an image layer (cf. Fig. 9). It is based on the assumption, that the used images include an infrared (IR) channel and are generated in summer, when the vegetation is in an advanced period of growth. The use of vector data as prior knowledge plays an important role, which is represented in the semantic model with an additional GIS-layer (ATKIS DLMBasis, cf. Section 3): Field boundaries and wind erosion obstacles are exclusively located in the open landscape, thus, further investigations are focussed to this area. Additionally, the objects road, river and railway are introduced in the semantic model as field boundaries with a direct relation from the GIS-layer to the real world (i.e. a road is a field boundary). Of course, the underlying assumption is based on correct GIS-objects. Modelling of the GIS-objects in the geometry and material layer together with the image layer is not of interest, because they do not have to be extracted from the imagery; thus, the corresponding parts are represented with dashed lines in Fig. 9. Nevertheless, additionally extracted objects which are not yet included in the GIS database can be introduced at any time.

The *field* is divided in the semantic model into field boundary and field area in order to allow for different modelling in the layers. The field boundary is a 2D elongated vegetation boundary, which is formed as a straight line or edge in the image. The field area is a 2D vegetation region, which is a homogeneous region with a high NDVI (Normalised Difference Vegetation Index) value in the colour infrared (CIR) image. The *wind erosion obstacle* is divided in hedge and tree row due to different available information from the GIS-layer, which is partially stored in the database. The wind erosion obstacles are not only described by their direct appearance in geometry and material, but also through the fact, that due to their height (3D object) there is a 2D elongated shadow region next to the object and in a known direction. In particular, the relationships between the objects to be extracted are of interest leading to connections within the layers: One object can be part of another one or be parallel and nearby, and together they form a context network in the real world. For instance, wind erosion obstacles are not located in the middle of a field because of disadvantageous cultivation conditions, but solely on the field boundaries.

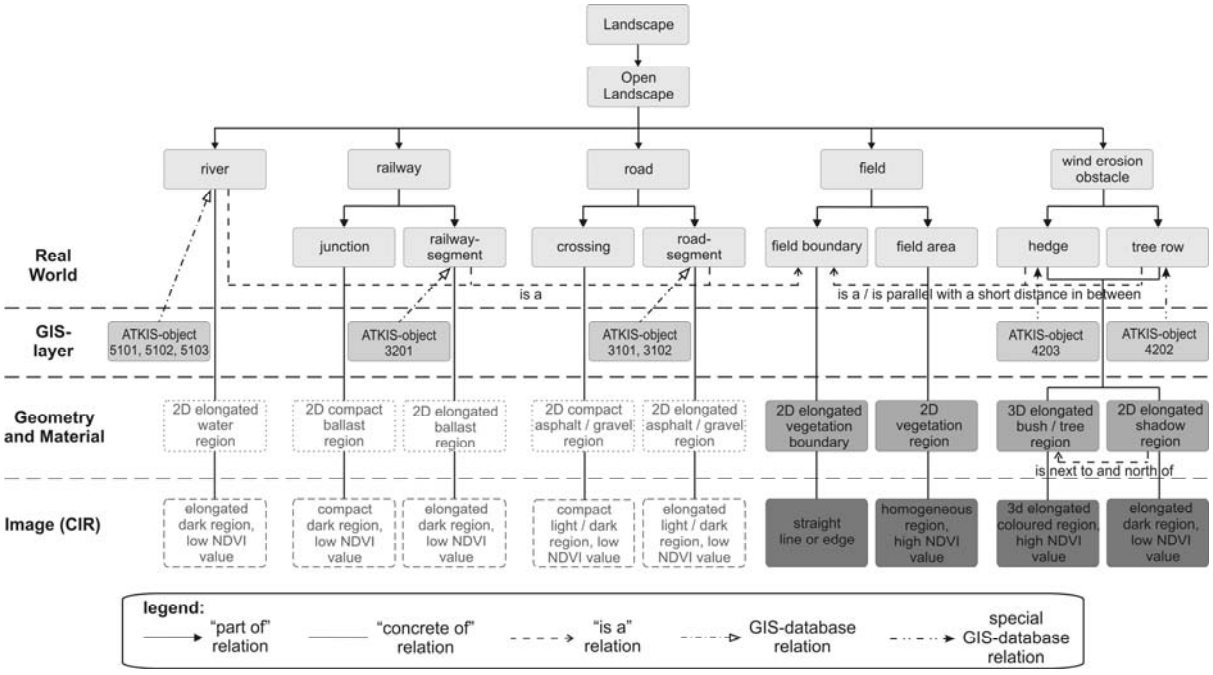


Fig. 9. Semantic Model.

The strategy derived from the modelled characteristics of the field boundaries and wind erosion obstacles aims at realising an automatic processing flow. Imagery and GIS-data are the input data to initialise the process: First, field boundaries and wind erosion obstacles are extracted separately. At the end, a combined evaluation of the preliminary results is advantageous due to the modelled geometrical and thematic similarities of the objects of interest getting a refined and integrated solution.
The strategy extracting the *field boundaries* starts with the derivation of the open landscape from the GIS data. In addition, within the open landscape, regions of interest are selected using the roads, rivers and railways as borderlines (cf. Section 4.2.2.). Consequently, the borderlines of the regions of interest are field boundaries, which are already fixed. In each region of interest a segmentation is carried out in a coarse scale ignoring small disturbing structures and thus exploiting the relative homogeneity of each field. The aim is to obtain a topologically correct result, even though the geometrical correctness may not be very high. Afterwards, network snakes are used to improve the preliminary field boundaries.
The strategy extracting the *wind erosion obstacles* starts again within the open landscape due to the modelled characteristics. Search buffers can be defined around the GIS objects roads, rivers and railways, because tree rows or hedges are often located alongside these objects, and have to be verified using the imagery. In contrary, there is no prior information about the location of all other wind erosion obstacles, which can lie anywhere within the open landscape. In addition to the modelled material characteristics, the geometrical part such as the straightness or minimum length has to be considered.
Finally, the combined evaluation of the preliminary results identifies discrepancies between the field boundaries and wind erosion obstacles. For example, extracted wind erosion obstacles without a corresponding extracted field boundary have to be checked, whether nearby a field boundary is missed, or whether the extraction of the wind erosion obstacle is wrong. Consequently, the combined evaluation and refined extraction process leads to a consistent and integrative final result.

4.2.2 Preparation of GIS data

As described in the previous section the regions of interest are primarily derived from roads, rivers and railways according to the GIS data, as far as these objects are located in the open landscape. The fact that these network generating objects imply a segmentation of the open landscape is used as starting point, as all necessary borderlines are already present in this data set – however, this segmentation is too extensive (e.g. because of administrative reasons). In order to detect only the borderlines concerning regions of interest, a topological data model is generated (Egenhofer et al., 1989), consisting of a embedded graph structure, which contains the open landscape regions and the network generating objects.
This graph based data model represents boundaries of area objects and one-dimensional objects as edges and therefore allows deciding if certain segmentations are a result of the separating network. The removal of all edges, which are not caused by the separating network, implies the merging of the adjacent regions and finally results in the generation of the regions of interest (Fig. 10 a).

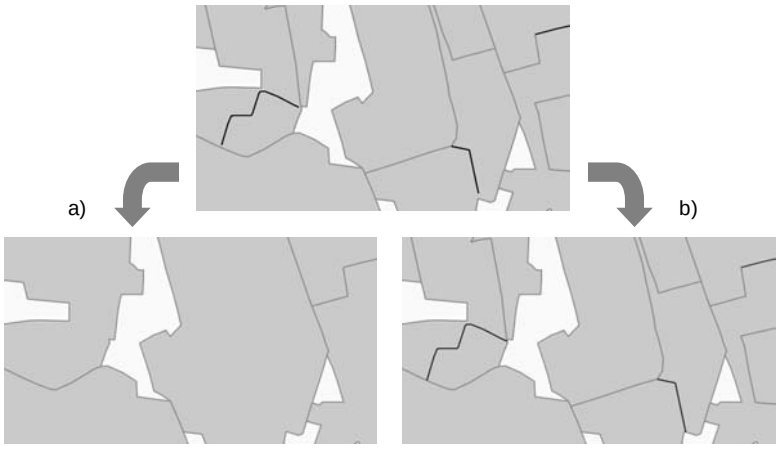


Fig. 10. Generation of regions of interest (a) and adjustment of tree rows and hedges (b).

Furthermore, the initial topological data model (i.e. before edge removal) is used to prepare the tree row and hedge objects used in the wind erosion obstacle extraction process by extending them to the next respective edge (Fig. 10 b, dark lines). This process of alignment is similar to topological error correction of inaccurately produced maps (Ubeda et al., 1997), whereas these objects are not inaccurately acquired but tree rows or hedges often end with a short distance to roads, rivers or railways.

4.2.3 Extraction of field boundaries

The extraction of field boundaries starts with a segmentation within each region of interest exploiting the modelled similar characteristics of each field. The border area of each region is masked out due to disturbing heterogeneities, which are typical for fields and deteriorate the subsequent steps. A multi-channel region growing is carried out using the RGB- and IR-channels of the images with a resolution of few meters. The four channels give rise to a 4-dimensional feature vector: Neighbouring pixels are aggregated into the same field region, if the difference of their feature vectors does not exceed a predefined threshold. In concert with the modelled constraints, the resulting field regions must have a minimum size. The case of identical vegetation of neighbouring fields may lead to missing boundaries. In order to overcome this problem, the standard deviation of the grey values in the image within a quadratic mask is computed, i.e. high values typically belong to field boundaries. Extracted lines from the standard deviation image within sufficiently large field regions are evaluated concerning length and straightness. Positively evaluated lines are used to split the initially generated field regions.

The result of the segmentation leads to topologically correct but geometrical inaccurate results. Network snakes are used to improve the geometrical correctness of the preliminary field boundaries while maintaining the topological constraints. Snakes were originally introduced in Kass et al. (1988) as a mid-level image analysis algorithm, which combines geometric and/or topologic constraints with the extraction of low-level features from images. A traditional snake is a parametric curve (Kass et al., 1988; Butenuth and Heipke, 2005)

$$v(s,t) = \left(x(s,t), y(s,t)\right) , \tag{2}$$

where s is the arc length, t the time, and x and y are the image coordinates of the 2D-curve. The image energy is defined as

$$E_I(v) = -\frac{1}{|v|}\int_0^{|v|} \left|\nabla I(v(s,t))\right| ds \ , \tag{3}$$

where I represents the image, $|\nabla I(v(s,t))|$ is the norm of the gradient magnitude of the image at the coordinates $x(s)$ and $y(s)$ and $|v|$ is the total length of v. In practice, the image energy $E_I(v)$ is computed by integrating the values $|\nabla I(v(s,t))|$ in precomputed gradient magnitude images along the line segments that connect the polygon vertices. The internal energy is defined as

$$E_{v(s,t)} = \tfrac{1}{2}\left(\alpha(s)\cdot\left|v'(s,t)\right|^2 + \beta(s)\cdot\left|v''(s,t)\right|^2\right), \tag{4}$$

where the function $\alpha(s)$ controls the first-order term of the internal energy: the elasticity. Large values of $\alpha(s)$ let the contour become very straight between two points. The function $\beta(s)$ controls the second-order term: the rigidity. Large values of $\beta(s)$ let the contour become smooth, small values allow the generation of corners. $\alpha(s)$ and $\beta(s)$ need to be predefined based on experimental data and experience.

The total energy of the snake, to be minimised, is defined as $E_{snake} = E_{v(s,t)} + E_I(v)$. A minimum of the total energy can be derived by embedding the curve in a virtual viscous medium solving the equation

$$\frac{\partial E_{v(s,t)}}{\partial v(s,t)} + \kappa\frac{\partial E_I(v)}{\partial v(s,t)} + \gamma\frac{dv(s,t)}{dt} = 0 \ , \tag{5}$$

where γ is the viscosity of the medium and κ is the weight between internal and image energy. After substituting of

$$\frac{\partial E_{v(s,t)}}{\partial v(s,t)} = A_{\alpha(s),\beta(s)}v(s,t) \quad \text{and} \quad \frac{dv(s,t)}{dt} = \left(v(s,t) - v(s,t-1)\right) \tag{6}$$

in equation 5, a solution for the contour at time t depending on time t-1can be computed:

$$V_{s,t} = \left(A + \gamma I\right)^{-1}\gamma V_{s,t-1} - \kappa\frac{\partial E_I(v)}{\partial v(s,t-1)} \ , (I\text{: identity matrix}) \tag{7}$$

$V_{s,t}$ stands for either X or Y, the vectors of the x and y coordinates of the contour. A is a pentadiagonal matrix, which depends only on the functions $\alpha(s)$ and $\beta(s)$.

A main problem of snakes is the necessity to have an initialisation close to the true solution. Methods to increase the capture range of the image forces are not useful in our case, because there are lots of disturbing structures within the fields, which can cause an unwanted image energy and therefore a wrong result. Thus, only the local image information is of interest. As described above, the result of the segmentation is used to initialise the processing.

In addition to the good initialisation the derivation of the *topology* of the initial contours is most important. The global framework of the accomplished segmentation gives rise to a network of the preliminary field boundaries: Enhancing traditional snakes, network snakes are linked to each other in the *nodal points* and thus interact during processing (cf. Fig. 13 b). Similarly, the connection of the *end points* of the contours to the borders of the region of interest must be taken into account: In contrast to the nodal points, a movement of the end points is only allowed along the borders of the regions of interest. These topological constraints are considered, when filling the matrix A (see equation 7) with the functions $\alpha(s)$ and $\beta(s)$, which in our case are taken to be constant.

4.2.4 Extraction of wind erosion obstacles

The extraction of wind erosion obstacles is concentrated to the open landscape, as pointed out in the semantic model. No other GIS data (road, river, railway) is used, i.e. no prior geometric information reduces the search area, in order to acquire all tree rows and hedges within the open landscape. A texture segmentation is accomplished in the CIR images with a resolution of few meters yielding the texture classes *tree/hedge*, *settlement area* and *agricultural area*, for details concerning the approach cf. Gimel´farb (1996). The training images are generated manually by a human operator. The texture class of interest *tree/hedge* is, as expected, fragmented and not complete. Therefore, the elongated and small regions of the class are vectorised. Starting point are the left and right boundaries of these regions: centrelines are then computed, which are evaluated concerning length and straightness. Currently, the third dimension, as described in the semantic model, is not used to extract the wind erosion obstacles due to a missing digital surface model.

5. Results

In this section some results of test areas in northern Germany are presented. First, results are shown to reveal the possibility to perform the alignment and change detection for the updating of vector data sets with a high degree of automation. Second, results of the extraction of field boundaries and wind erosion obstacles are highlighted to demonstrate the capability of the described methods.

5.1 Results of the vector data integration

In Fig. 11 the results of the different alignment methods can be seen. The ICP algorithm using an iterative four-parameter transformation is very suitable for the alignment of objects which already have a similar geometry. The alignment parameters which are the results of the ICP algorithm can give a first hint whether the geometric discrepancies are due to map creation and acquisition methods (a., d.) or to changes which occurred to the real world object (c.). The resulting scale factor which was calculated for the alignment of object c. was rated as too large and therefore no alignment was performed. Of course changes of the topography can not be discovered by simple evaluation of these parameters. For object b. the algorithm achieved a best fit with four parameters below certain thresholds, but the remaining differences between the geometries still have to be corrected.

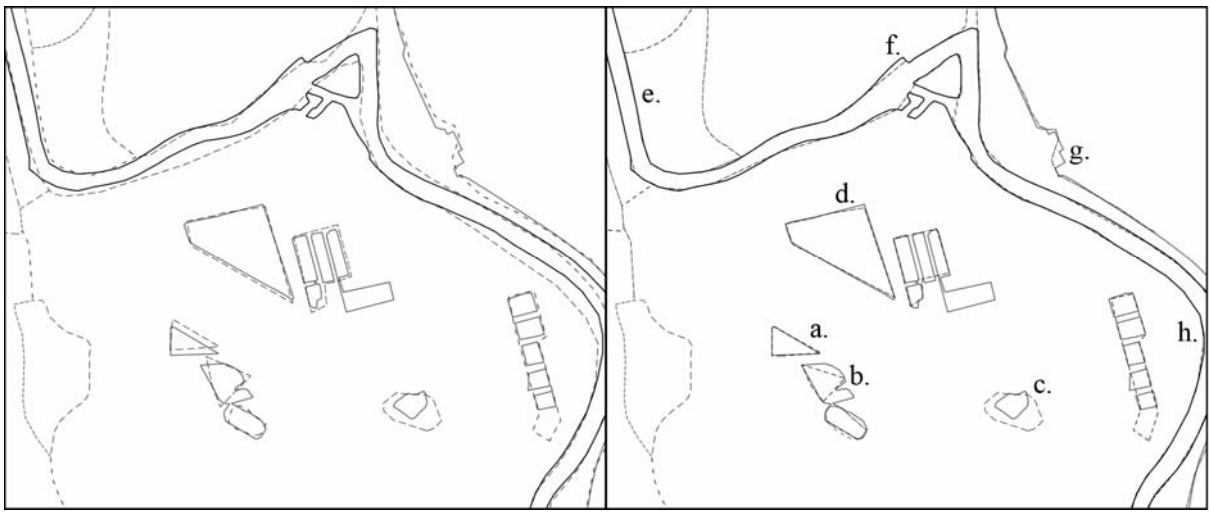


Fig. 11. Result of the approach, GK 25 (thin, dotted line) aligned on the reference German digital topographic map (ATKIS, dark lines).

The DIA implementation showed very good results to compensate local discrepancies which can not be corrected using the four-parameter-ICP, as it aims for the best alignment of the whole object. There is no single four-parameter transformation which is capable of adjusting large extended natural objects like rivers object to their corresponding partners in ATKIS, so that parts e, f and h would be properly aligned. The results does exhibit some small gaps between the geometries (see e.g. area (g)): this is due to the fact that the DIA algorithm in the current version is only working only with existing vertices without inserting additional ones.

In order to identify possible changes between the objects in the different representations, after the alignment process has been completed an intersection of corresponding objects is used for the change detection. The intersection is performed on all types of topographic elements which are represented in the data sets. The results of the intersection process will be evaluated and according to their semantic attribution sorted into three different classes.

- *Type I* : Segment has corresponding semantic attribute in both data-sets, no adaptation required,
- *Type II* : Segment has different semantic attributes, and a suitable information can be derived from the reference data set and the candidate data set will be updated.
- *Type III* : Segment has different semantic attributes, but a suitable information can not be derived from the reference data set. Manual interaction is required.

Type II will also be assigned to objects which are represented in the reference, but not the candidate data-set, this is the result of different updating periods between the reference and the candidate data set, which result in outdated objects. While *Type I* and *II* require only geometric corrections or attribute adaptation and can be handled automatically, *Type III* needs more of the operators attention. Depending on the size and the shape of a *Type III* segment and by using a user-defined threshold, these segments can be filtered, removed and the remaining gap can be corrected automatically, this will avoid the integration of sliver polygons and segments which are only the results of geometric discrepancies.

5.2 Results of the raster and vector data integration

5.2.1 Results of the extraction of field boundaries

Results of the proposed strategy to extract field boundaries are presented in this section. The result of the first step, the segmentation, is shown in Fig. 12: The boundaries of the regions of interest are depicted in black, the preliminary field boundaries are depicted in white. Compared to reference data, the completeness of the segmentation within a test area of 25 km^2 (Lower Saxony, North of Germany) is 73 %, the correctness is 82 % and the rms error computed by considering the horizontal derivation between extracted and reference result is 5.8 m or 3 pixels. The quality of the results is promising, but as expected the geometrical correctness is not very high.
One region of interest is selected to demonstrate the methodology of the network snakes (cf. Fig. 13 a-d): The initialisation of the snake – equivalent to the result of the segmentation – is shown in the first figure. The topology is pointed out in Fig. 13 b): The individual snakes forming the network are linked to each other in the nodal point (black), and the end points (black with white hole) are linked to the boundary of the region of interest. The movement of the snake superimposed to the standard deviation image is shown in Fig. 13 c), the final result superimposed to the real image in Fig. 13 d). The example demonstrates, that network snakes are a useful tool to improve the geometrical correctness of topologically correct but geometrically inaccurate results.

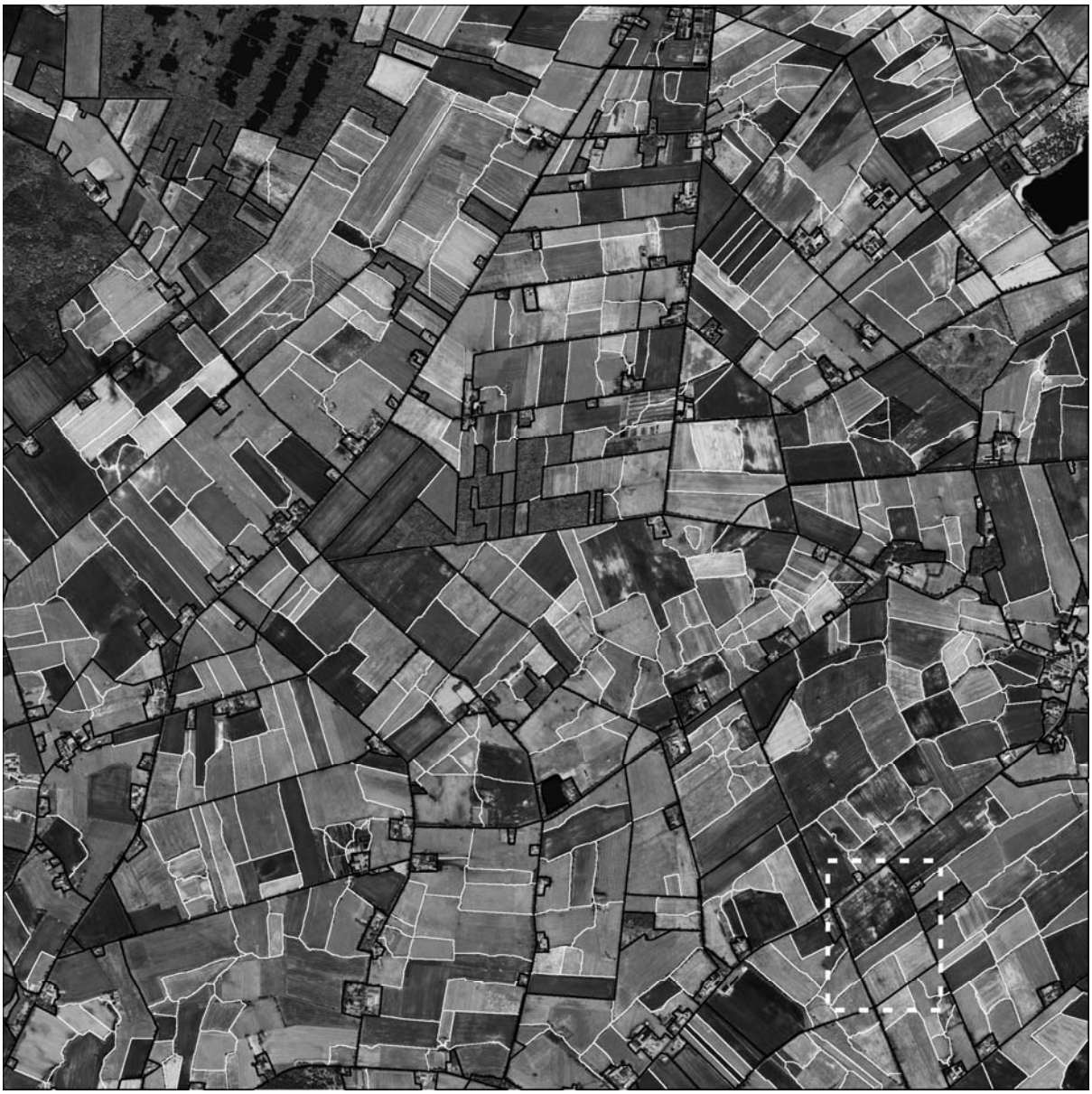
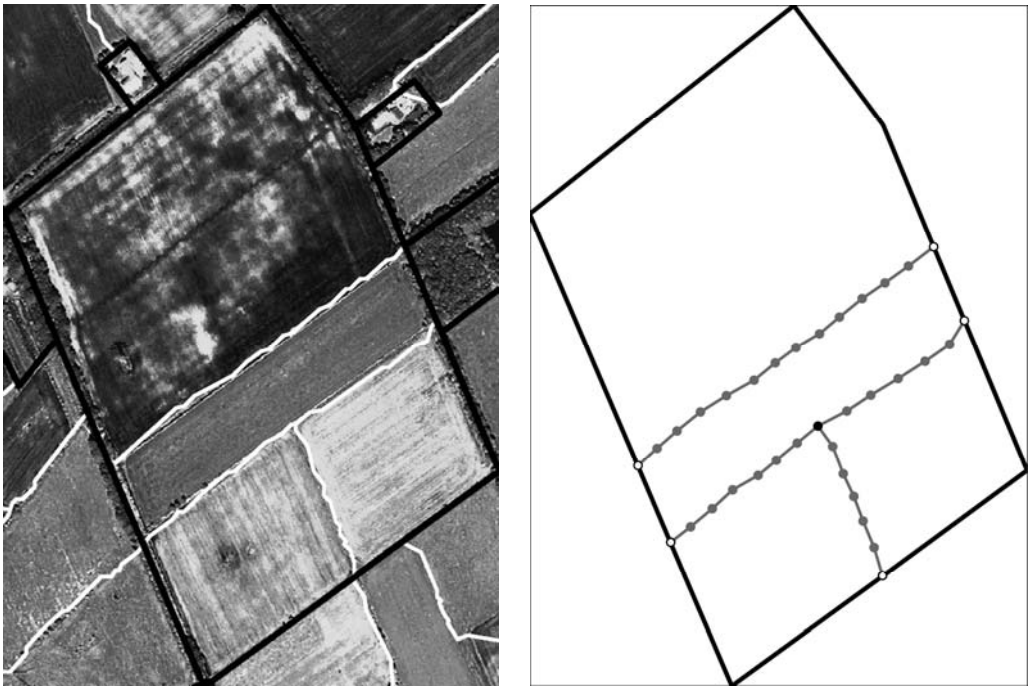
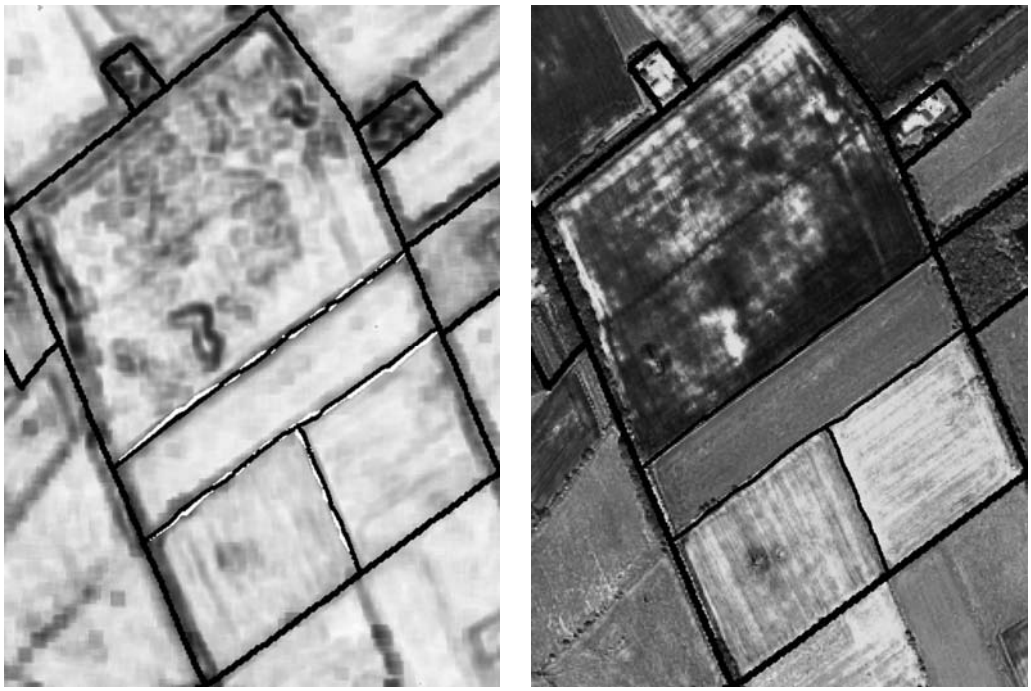


Fig. 12. Result of the segmentation.



a) b)



c) d)

Fig. 13. Results of the use of network snakes: a) initialisation, b) building the topology, c) initialisation (white) and movement of the snake (black), d) extracted field boundaries.

5.2.2 Results of the extraction of wind erosion obstacles

The result of the texture segmentation is presented in Fig. 14: The class *tree/hedge* is depicted in white, the class *agricultural area* in light grey and the class *settlement area* in dark grey. The parts of the image, which do not belong to the open landscape exploiting the prior GIS knowledge, are marked in black. The fragmented class *tree/hedge* is vectorised yielding the wind erosion obstacles, as depicted in Fig. 15 in white. The texture segmentation works well, but an additional digital surface model is needed to improve and stabilise the results.

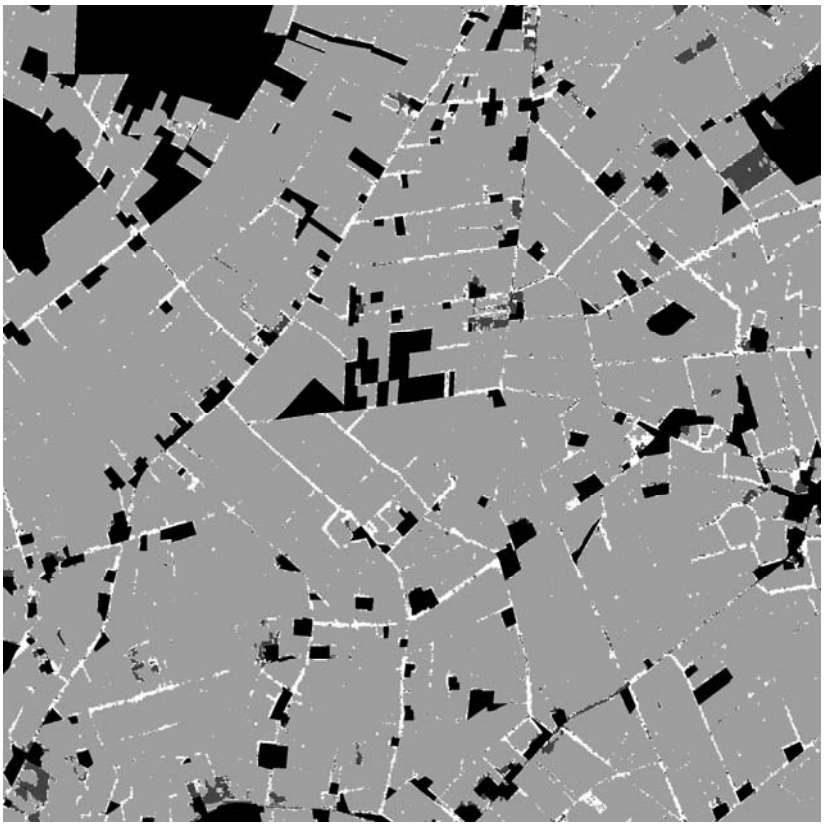


Fig. 14. Result of the texture segmentation.

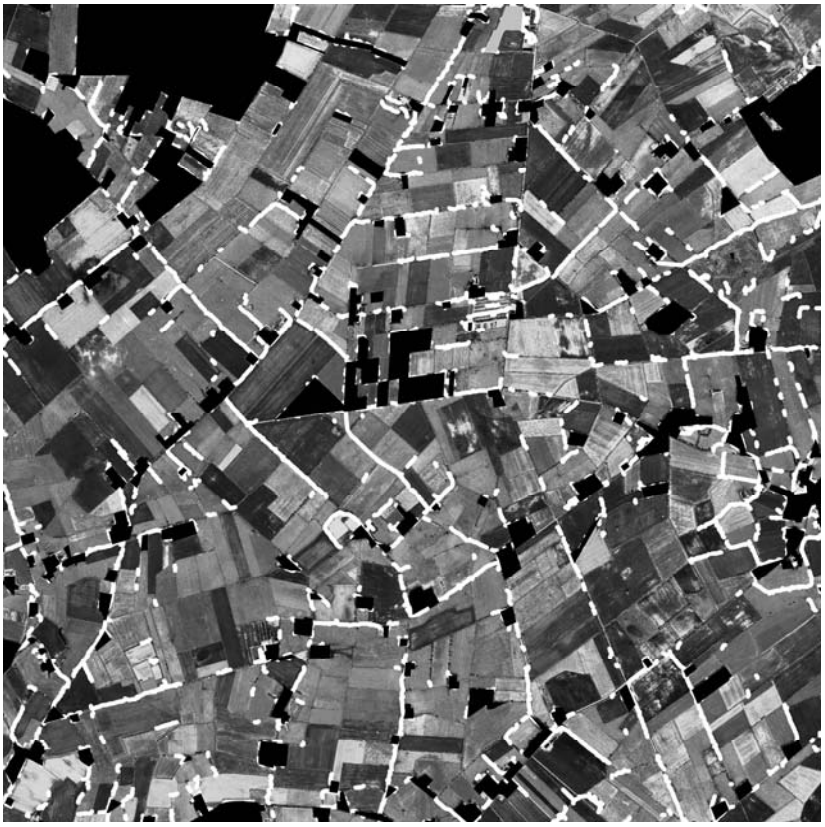


Fig. 15. Result of the extracted wind erosion obstacles.

6. Conclusions

The geospatial federated database has provided the expected access to the involved data sets and to the results of the matching and extraction processes. It provides a basis not only for querying linked objects but also for update propagation. Utilised appropriate data structures like topological data models offer further approaches to assure topological consistence during geometric alignment and to accomplish a structural graph based matching.
The geometric comparison and the derivation of object links, together with the ICP and DIA alignment followed by rubber-sheeting and the evaluation process show good results. So far this strategy was used with one data set as reference which remains unchanged, while a second data set is adjusted, but it can also be adapted to other vector based conflation tasks requiring an intermediate geometry. Depending on the selected thresholds large discrepancies of the shape boundaries can consider as outliers and can be treated accordingly in the subsequent overlay and analysis step. While matching can be performed automatically, there are still some steps during geometric alignment and change detection which require the decision of a human operator, but the high degree of automation reduces the manual process considerable. Future work will concentrate on developing a strategy to also automate these processes. Especially the selection of the appropriate alignment method and the corresponding thresholds will be enhanced.
The method integrating raster and vector data by means of the extraction of field boundaries and wind erosion obstacles from imagery exploiting prior GIS knowledge has also chosen promising results. Concerning the extraction of field boundaries the basic step of the strategy, the segmentation, could be enhanced by using an additional texture channel to prevent wrong field boundaries. They occur, when there are large heterogeneities within a field. The control of the network snakes could be improved by selecting variable values when filling the matrix A to increase the geometrical correctness, the use of network snakes provide a topologically consistent solution. Regarding the extraction of wind erosion obstacles, initial results show the potential, but also the limitations of the current approach. The use of a digital surface model will probably be very helpful to achieve better results. Finally, the combined evaluation of the field boundaries and wind erosion obstacles will identify discrepancies between the different extracted objects, resulting in a more consistent and integrative final result.

References

Baltsavias, E.P. (2004): Object Extraction and Revision by Image Analysis Using Existing Geodata and Knowledge: Current Status and Steps towards Operational Systems. ISPRS Journal of Photogrammetry and Remote Sensing 58 (3-4), 129-151.

Batini, C., Lenzerini, M., Navathe, S. B. (1986): A Comparative Analysis of Methodologies for Database Schema Integration. ACM Comput. Surv. 18(4), 323-364.

Beeri, C., Doytsher, Y., Kanza, Y., Safra, E., Sagiv, Y. (2005): Finding Corresponding Objects when Integrating Several Geo-Spatial Datasets. Proc. 13th ACM International Symposium on Advances in Geographic Information Systems, Bremen, Germany, 4-5 November 2005, 87-96.

Besl, P., McKay, N. (1992): A Method for Registration of 3-D Shapes, IEEE Transactions on Pattern Analysis and Machine Intelligence (Special issue on interpretation of 3-D scenes - part II) 14 (2), 239-256.

Butenuth, M. (2004): Modelling the Extraction of Field Boundaries and Wind Erosion Obstacles from Aerial Imagery. International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences XXXV (Part B4), 1065-1070.

Butenuth, M., Heipke, C. (2005): Network Snakes-Supported Extraction of Field Boundaries from Imagery. In: Kropatsch, Sablatnig, Hanbury (Eds), 27th DAGM Symposium, Wien, Österreich, Springer LNCS 3663, 417-424.

Conrad, S. (1997): Föderierte Datenbanksysteme, Springer-Verlag, Berlin.

Devogele, T., Parent, C., Spaccapietra, S. (1998): On spatial database integration, International Journal of Geographical Information Science, 12:4, 335-352.

Doytsher, Y. (2000): A rubber sheeting algorithm for non-rectangular maps, Computer & Geosciences, 26 (9-10), 1001-1010.

Egenhofer, M.J., Frank, A.U., Jackson, J.P. (1989): A Topological Data Model for Spatial Databases, Lecture Notes in Computer Science 409, 271-286.

Gimel´farb, G.L. (1996): Texture modelling by multiple pairwise pixel interactions. IEEE Transactions on Pattern Analysis and Machine Intelligence 18 (11), 1110-1114.

Goesseln, G. v., Sester, M. (2004): Integration of geoscientific data sets and the german digital map using a matching approach. International Archives of Photogrammetry and Remote Sensing 35 (Part 4B), 1249-1254.

Hettwer, J., Benning, W. (2000): Nachbarschaftstreue Koordinatenberechnung in der Kartenhomogenisierung, Allg. Verm. Nachr. 107, 194-197.

Hill, D. A. and Leckie, D. G. (Eds.) (1999): International forum: Automated interpretation of high spatial resolution digital imagery for forestry. February 10-12, 1998, Natural Resources Canada, Canadian Forest Service, Pacific Forestry Centre, Victoria, British Columbia.

Kass, M., Witkin, A., Terzopoulus, D. (1988): Snakes: Active Contour Models. International Journal of Computer Vision 1, 321-331.

Kohonen, T. (1997); Self-Organizing Maps. Springer.

Kleiner, C., Lipeck, U., Falke, S. (2000): Objekt-Relationale Datenbanken zur Verwaltung von ATKIS-Daten. In: Bill, R., Schmidt, F.: ATKIS - Stand und Fortführung, Verlag Konrad Wittwer, Stuttgart, 169-177.

Laurini, R. (1998): Spatial multi-database topological continuity and indexing: A step towards seamless GIS data interoperability, International Journal Geographical Information Science, 12:4, 373-402.

Löcherbach, T. (1998): Fusing Raster- and Vector-Data with Applications to Land-Use Mapping. Inaugural-Dissertation der Hohen Landwirtschaftlichen Fakultät der Universität Bonn.

Mantel, D., Lipeck, U. W. (2004): Datenbankgestütztes Matching von Kartenobjekten. In: Arbeitsgruppe Automation in der Kartographie - Tagung Erfurt 2003, BKG, Frankfurt, 145-153.

Mayer, H. (1999): Automatic Object Extraction from Aerial Imagery - A Survey Focusing on Buildings. Computer Vision and Image Understanding 74 (2), 138-149.

Rigeaux, P., Scholl, M., Voisard, A. (2002): Spatial Databases with Application to GIS, Morgan Kaufman Publishers.

Sattler, K.-U., Conrad, S., Saake, G. (2000): Adding Conflict Resolution Features to a Query Language for Database Federations, 41-52.

Sester, M., Hild, H. & Fritsch, D. (1998): Definition of Ground-Control Features for Image Registration using GIS-Data. In: Schenk, T. & Habib, A. (Eds.), IAPRS 32/3, ISPRS Commission III Symposium on Object Recognition and Scene Classification from Multispectral and Multisensor Pixels, Columbus/Ohio, USA, 537-543.

Torre, M., Radeva, P. (2000): Agricultural Field Extraction from Aerial Images Using a Region Competition Algorithm. International Archives of Photogrammetry and Remote Sensing XXXIII (Part B2), 889-896.

Yuan, T., Tao, C. (1999): Development of conflation components. In: Li, B., et al. (Eds.), Geoinformatics and Socioinformatics – The Proceedings of Geoinformatics ´99 Conference, Ann Arbor, USA, 19-21 June 1999, 1-13.

Walter, V. & Fritsch, D. (1999): Matching Spatial Data sets: a Statistical Approach, International Journal of Geographical Information Science 13(5), 445–473.

Ubeda, T., Egenhofer, M. J. (1997): Topological Error Correcting in GIS. In: Advances in Spatial Databases, 5th International Symposium, SSD'97.