D. Fritsch, M. Englich & M. Sester, eds, 'IAPRS', Vol. 32/4, ISPRS Commission IV Symposium on GIS - Between Visions and Applications, Stuttgart, Germany.

CONTRIBUTIONS TO A QUALITY DESCRIPTION OF AREAL OBJECTS IN SPATIAL DATA SETS

Lemonia Ragia
Institute of Photogrammetry
University of Bonn, Nußallee 15
D-53115 Bonn, Germany
Ph.: +49-228-732968
Fax: +49-228-732712
Lemonia.Ragia@ipb.uni-bonn.de

Stephan Winter
Department of Geoinformation
TU Vienna, Gusshausstr. 27-29
A-1040 Vienna, Austria
Ph.: +43-1-588013788
Fax: +43-1-5043535
winter@geoinfo.tuwien.ac.at

KEY WORDS: spatial data quality, comparison of spatial entities, building extraction.

ABSTRACT

In this paper we present a quality evaluation of two-dimensional building acquisition. We propose methods for identification and quantification of differences between independently acquired regions, and we present a systematic classification of those differences.
Differences between acquired sets $\mathcal{R}_j = \{r_i\}_j$ of regions r_{ij} depend on the context of observation, on the technique of observation, and so on. We distinguish *topological* and *geometrical* differences. Topological differences refer to the interior structure of a set of regions as well as to the structure of the boundary of a single region. Geometrical differences refer to the location of the boundary of a single region or of a set of regions, independent of their representation and of the structure of the boundaries.
Identification of differences requires a matching of two data sets $\mathcal{R}_1$ and $\mathcal{R}_2$ which is done here by weighted topological relationships. For the identification of topological differences between two sets $\mathcal{R}_1$ and $\mathcal{R}_2$ of regions we use the two region adjacency graphs. For an identification of geometrical differences we use the zone skeleton between two matched subsets r_p1 and r_q2 of the given sets. The zone skeleton is labeled with the local distances of the corresponding boundaries of the subsets; especially we investigate its density function.
An example, based on two real data sets of acquired ground plans of buildings, shows the feasibility of the approach.

1 INTRODUCTION

1.1 Motivation and Idea

A lot of concurrent acquisition techniques for spatial databases are available. They need to be evaluated with respect to given specifications, and compared with respect to a usually great number of criteria, such as efficiency, accuracy, and completeness. In this paper we present a quality evaluation of two-dimensional building acquisition. We propose methods for identification and quantification of differences between independently acquired regions, and we present a systematic classification of those differences.

Two data sets of the same object category nearly never show incident entities (Fig. 1). Differences between acquired sets $\mathcal{R}_j = \{r_i\}_j$ of regions r_{ij} depend on the *context* of observation, on the *technique* of observation, on *error*, on numerical problems of *discrete representations*, on *temporal effects*, and so on. When given a set of observed regions a second set of regions is needed, playing the role of a reference data set. Fig. 1 shows two small sets of ground plans of buildings acquired with two different techniques. Neither the number of buildings nor the partitioning coincides. Additionally one could imagine different levels of detail in the data acquisition, and there will be differences in the location of matching regions. The *identification* of the various differences obviously is a complex problem as all types of errors may interfere.

In this paper we investigate the differences between two independently acquired data sets which cover the same two-dimensional space, with one data set considered as the reference. We distinguish *topological* and *geometrical* differences. Topological differences refer to the interior structure of a set of regions as well as to the structure of the boundary of a single region. Geometrical differences refer to the location of boundaries of single regions or of sets of regions, independent of their representation and of the structure of their boundaries. By this way we obtain a complete description of the differences of two sets of regions.

Identification of differences requires a matching of two data sets $\mathcal{R}_1$ and $\mathcal{R}_2$ which is done here based on weighted topological relationships. For the identification of topological differences between two sets of regions, $\mathcal{R}_1$ and $\mathcal{R}_2$, we use the two region adjacency graphs. For an identification of geometrical differences we use the zone skeleton between two matched subsets r_{p1} and r_{q2} of the given sets. The zone skeleton is labeled with the local distances of the corresponding boundaries of the subsets; especially we investigate its density function.

Ultimate goal of our work are tools for determining the quality of sets of acquired two dimensional regions. The use of such a description of differences is manifold:

- Data capture can be supplied by classification of errors on-the-fly.
- Comparison of two data sets yields measures for quality assessment.
- Having a complete statistics of classified differences may be advantageous when merging two data sets.

1.2 Framework and Terms of the Paper

In the context of this paper we confine ourselves to data sets containing the same class of areal objects (regions). We do not consider points and lines as spatial objects; objects in real world have spatial extent, and they need a representation adequate for their extent. Also we exclude here fields (coverages) as unbounded spatial phenomena (Burrough and Frank, 1996). – The objects shall belong to the same class because they will be compared only with regard to their topology and geometry, but not in their attributes or other properties. That presumes also a similar granularity in both data sets; it does not make sense to compare, e.g., buildings with street-blocks.

1.3 Structure of the Paper

The paper is organized as follows. Section 2 summarizes the previous work in the field of quality descriptions for spatial data, and in comparing spatial data sets. In Section 3 we investigate differences between spatial entities from a topological and a geometrical point of view, and we discuss the reasons for observable

D. Fritsch, M. Englich & M. Sester, eds, 'IAPRS', Vol. 32/4, ISPRS Commission IV Symposium on GIS - Between Visions and Applications, Stuttgart, Germany.

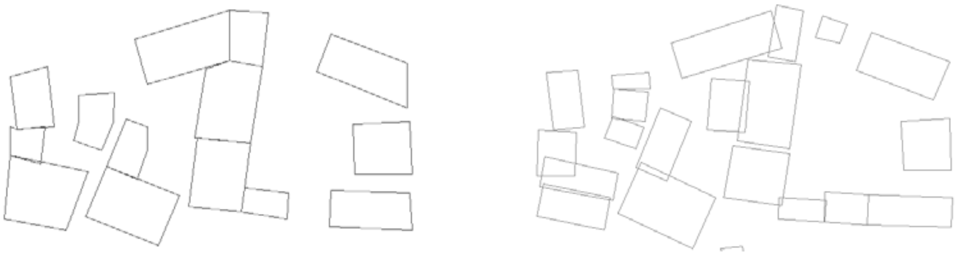

Figure 1: Two real data sets $\mathcal{R}_1$ and $\mathcal{R}_2$, observed with different acquisition techniques (©DeTeMobil 1998).

differences. Section 4 presents the used approaches for characterizing topological and geometrical differences of regions. With these tools at hand, in Section 5 we investigate and systematically classify geometric differences. That will yield the basic parameters to find operational quality descriptions. In Section 6 we present an empirical test of the method. We conclude in Section 7 with an overview on further investigations and developments in this field.

2 PREVIOUS WORK

In this section we discuss previous spatial data set comparisons and quality descriptions. It will become clear that there is a strong demand for further research; the presented ideas fit into this framework.

2.1 Comparison of Spatial Data Sets

Comparison of two spatial data sets, referring to the same section of space, is a fundamental ability for Geographic Information Systems (GIS); it is utilized in handling positional uncertainty or imprecision (Glemser, 1993), in matching spatial entities (Knorr et al., 1997, Harvey et al., 1998), in data fusion (Haala, 1994), in change detection, in generalization (Goodchild and Proctor, 1997, Tryfona and Egenhofer, 1997), and so on. Comparison yields a description based on common and distinct features (Tversky, 1977); the type of description – number and meaning of parameters – depends on the context.

The presented ideas are related to actual work in Maine (Bruns and Egenhofer, 1996, Egenhofer et al., 1997). They describe spatial *scene similarity* by distances taken from conceptual neighborhood graphs of different qualitative relations between regions. Such a *distance* is discrete and qualitative; as a weighted mean over the different relations it is abstract and cannot be interpreted geometrically. In contrast (and complementing) we use a continuous measure: the *distance function* between the two boundaries (Winter, 1996). The distance function can be interpreted geometrically, and additionally we are able to categorize the results for qualitative distinctions. Additionally, we compare the internal topological structure of the associated regions, which complements also a comparison of relations.

2.2 Spatial Data Quality Descriptions

Quality of spatial data (sets) is an actual topic in standardisation as well as in research. The CEN metadata standard (Comité Européen de la Normalisation), containing also spatial data quality descriptions, is pending for resolution, and also ISO TC 211 is finalizing with a standard. Standards provide a common set of parameters to describe a phenomenon, here spatial data quality, which is necessary for data cataloguing as well as for data exchange. The actual need for standards is beyond doubt, for reasons of market development. But the process of specification took several years, and the difficulty to set standards in the field of spatial data quality indicates that theory is not well developed. Research started with collections of quality aspects (Guptill and Morrison, 1995, Aalders, 1996), which influenced directly standardisation. A test with the pending CEN standard concluded with critique on practical applicability, especially from the users' point of view (Timpf et al., 1996). A theoretical framework for spatial data quality is lacking up to now.

In absence of such a framework investigations of spatial data quality will lead to context-specific methods. In principle, one has to reference (a) to ground truth, (b) to a reference data set, (c) to a large number of other data sets, or (d) to quality descriptions extracted from knowledge about data capture and experience (Baarda, 1967). In this paper we present a systematic analysis of differences between regions from two data sets, applying the approach (b). Based on this analysis we propose methods to detect and identify the differences. Without prior knowledge about the compared data sets, one is able to describe the differences between the data sets. If one data set can be used as reference, the statements can be turned into quality descriptions.

3 INVESTIGATION OF THE PROBLEM

In this section we specify our task, discuss representational issues regarding comparisons, and categorize the differences that exist between matching regions from different data sets.

3.1 The Task

We assume that two sets $\mathcal{R}_j$, $j = 1, 2$, of regions r are given: $\mathcal{R}_j = \{r_i\}_j$. Each region r_{ij} is described geometrically. The regions within one set may be related, e. g. by specifying a neighbourhood relation. These relationships are representable in a region adjacency graph (RAG) $G_j(\mathcal{R}_j, \mathcal{N}_j)$, consisting of the regions of $\mathcal{R}_j$ as nodes, and of the neighbourhood relations $\mathcal{N}_j$ as edges. For simplicity, we assume the geometry of the regions are described by polygons. This restriction does not result in conceptual limitations within our context. The task is to characterize the structural and geometrical differences between such two sets, and to develop tools for testing the equivalence of the two sets, or for detecting and identifying the differences, respectively.

3.2 Representation of Areal Objects

Spatial representations in GIS are classified as vector and raster, which are dual in the sense of space bounding and space filling (Peuquet, 1984). Both concepts describe location in a reference frame, e.g. in $\mathbb{R}^2$ or $\mathbb{Z}^2$. In a vector model, regions are represented by their boundary curves, which usually are regular closed polygons of *nodes* and *edges*. In contrast, in a raster model regions are represented by the set of raster elements which belong – at least to a dominant part of each element – to the object.

The shape of regions is independent from their representation, but the shape as well as the representation determine the kind and number of distinguishable topological relationships (Egenhofer et al., 1994, Winter, 1995). Shape may be arbitrarily complex: simply connected, multiple connected, i.e. with holes, or not connected, i.e. aggregations of disjunct parts. Additionally, in vector representation the inner geometry may be decomposed into parts. That may be caused by pure geometric reasons, like a decomposition into cell complexes, e.g. as in finite element methods, or by semantic reasons, when the object is aggregated from distinguishable parts. The latter often happens for man-made objects, like buildings (Gülch, 1997). The partition between two data sets may differ significantly.

It turns out that comparison of topology and of geometry requires a *scale* (Stevens, 1946) and a *common spatial representation*. The scale defines context and a measure, which is introduced both in the next section.

Determination of topological differences is performed easily using a vector representation. For the analysis, in Section 4.1 a mapping is proposed from geometric polygons to the dual, a *region adjacency graph*, which keeps only topological information.

In contrast, measures for differences in geometry will be based on vector properties as well as on a discrete space (Section 4.3). For that reason, geometry is investigated using a hybrid raster that preserves topological properties of $\mathbb{R}^2$. The hybrid raster representation consists of a decomposition of the plane $\mathbb{R}^2$ into uniform cell complexes, where the grid of $\mathbb{Z}^2$ is chosen as the set of 0-cells (or nodes), which are connected by axis-parallel 1-cells (or edges). Both together form an 1-skeleton, which encloses the 2-cells (or faces) (Kovalevsky, 1989, Winter, 1998a).

It is presumed that the data sets are topologically consistent. Such a property of a data set can be checked automatically in spatial databases (Joos, 1996).

Nevertheless, one has to define in detail what is topological consistence. Consider, e.g., a system for building extraction, based on volumetric primitives. An intersection of the primitives is possible and allowed (constructive solid geometry works with unification of elements). In the measuring process the operator adapts the form and pose parameters of the primitives. Complex buildings b are composed by some primitives p: $b = \bigcup_i p_i$. Even when buildings are touching intersections occur, guaranteeing to avoid small gaps between the buildings.

Especially qualitative methods of comparison have to be robust to treat effects from imprecise geometry. That is reached by tolerances, and by weighted topological relationships (Section 4.3).

3.3 Classification of Differences of Sets of Regions

Differences between sets of regions can be classified into *topological differences*, which are qualitative, and *geometrical differences*, which are quantitative. In Fig. 2 they are specified into inner structure, boundary structure, and location.

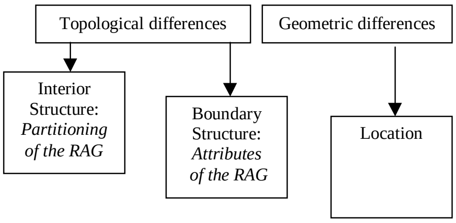
Topological differences
Geometric differences
Interior Structure: *Partitioning of the RAG*
Boundary Structure: *Attributes of the RAG*
Location

Figure 2: Classification of differences between two matched regions (*RAG*: region adjacency graph).

Differences in the Inner Structure Differences in the region adjacency graphs (Section 4.1) of two sets of regions require a mapping between the two graphs. This correspondence may be established by investigating the spatial relationships between all regions r_{i1} of the first set with all regions r_{i2} of the second set. We assume this mapping be established, e.g. by treating regions r_{i1} and r_{i2} as corresponding if they overlap strongly, coincide or contain each other (cf. T_2 in Eq. 2). In the most simple case we may identify four types of differences (cf. (Fuchs et al., 1994)):

- A region is missing.
- A region is spurious.
- Two or more regions are merged into one.
- A region is split into two or more regions.

In general, however, this mapping may be $n : m$, which results in a complex correspondence relation between the two sets of regions.

Differences in the Boundary Structure Assuming the boundaries of regions to be represented as polygons, or more general as possibly labeled sequence of boundary elements, we can identify differences between two matching structures without analysing geometry. Only in case of structural differences, we need to check the equivalence of the geometry of the boundaries. E.g., if a region has 4 points and the corresponding region has 5 points, the geometry may be identical, in case 4 points are identical and the other lies on one of the straight edges. In case the two structures are identical, a direct comparison of the attributes of the boundary elements may be sufficient, e.g. by just comparing the coordinates of corresponding polygon nodes.

This type of analysis can be applied also to the boundary of two corresponding *sets* of regions $\{r_k\}_1$ and $\{r_l\}_2$, assuming that the matching process has identified these two sets as corresponding. For example, if r_{a1} and r_{b1} from $\mathcal{R}_1$ are corresponding to r_{a2} from $\mathcal{R}_2$, one can determine the difference in the structure of the boundaries of $(r_{a1} \cup r_{b1})$ and r_{a2}.

Differences in Location The determination of differences in the location of two corresponding regions or two corresponding sets of regions requires a measure of similarity. This may be, e.g., the mean or the maximal distance (Hausdorff distance (Serra, 1982)) between corresponding points of the boundary. In order to be able to establish such a measure, the expected uncertainty of the region boundaries needs to be smaller than the regions themselves. As the geometric differences between two boundaries may be arbitrarily complex there is no canonical way to measure this difference. However, in case of quite similar and smooth boundaries, this distance may be easily established (Section 4.3).

Also here *sets* of regions may be compared. For example, a complex building may be partitioned into r_{a1} and r_{b1} in $\mathcal{R}_1$, and in r_{a2} and r_{b2} in $\mathcal{R}_2$. Matching should have identified r_{a1} corresponding to r_{a2}, and r_{b1} to r_{b2}, however $r_{a1} \ncong r_{a2}$ and $r_{b1} \ncong r_{b2}$. Then the common boundary of $r_{a1} \cup r_{b1}$ is to be compared to $r_{a2} \cup r_{b2}$. – The following differences are of special interest:

- two matched regions differ in single boundary points;
- two matched region differ in pose (shift, or rotation);
- two associated regions differ in form parameters.

4 USED TECHNIQUES

In this section the techniques are presented which are used later to characterize the discussed topological and geometric differences in two sets of regions. Especially we introduce the region adjacency graph, a refinement of topological relationships, and a local distance function with its histogram.

4.1 The Region Adjacency Graph

The *region adjacency graph* (RAG, Fig. 3) is a concept used in image processing. It assumes a complete partition of the plane. It is defined as the dual to the cell graph of connected regions.

In our context the RAG describes an arbitraty set of regions. The nodes of the RAG represent the components of the regions; their attributes are the number of nodes of the bounding polygon of the represented component. Two nodes of the RAG are connected by an edge if the components are in any topological relation different from *disjunct*. The attributes of the edges specify this topological relation; three types of relations are considered *touch*, *weak overlap*, and *strong overlap* (cf. Fig. 3). This set of relationships is not purely topological; there are metric influences, and groups of topological relationships are combined (the meaning of the notions differs from the next section):

- two regions touch if they have common parts in their boundaries, but do not overlap; they touch weakly if a remaining gap is lower than a tolerance;
- two regions overlap weakly if their intersection set is smaller than 50% of the smaller region (Eq. 1); and
- (in the context of the RAG:) strong overlap sums up all topological relations of the cluster T_2 (Eq. 2).

Consider Fig. 3 for an example. This figure describes the RAG of two sets of regions. The graph represents the right-hand part of the scene in Fig. 1, which corresponds to the image in Fig. 4. There are one complex and three single buildings in the first set of regions, and one complex and four single buildings in the second set of regions. In the first RAG, the number of boundary points of the components of the buildings is always four, and the relations between the components are *weak overlap* and *touch*. In the second RAG, the number of the boundary points varies, and the topological relation is only *touch*. The complex building consists of eight segments in the first set and of five components in the second set. There exist three single buildings in the first set as in the second one, but one of the first data set cannot be matched.

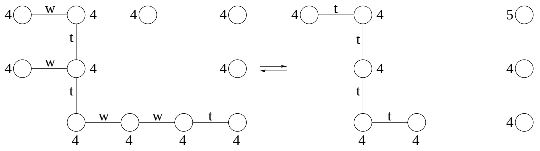

Figure 3: The region adjacency graph; *t*: *touch*, *w*: *weak overlap*, *s*: *strong overlap*.

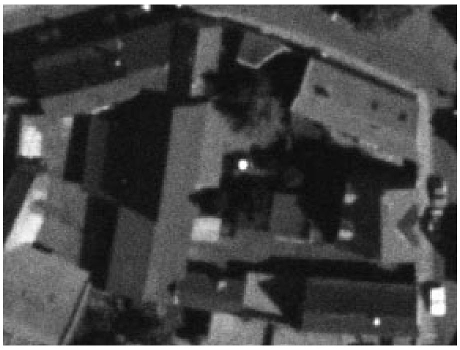

Figure 4: The corresponding image (©DeTeMobil 1998).

4.2 Topological Relations between Regions

In this section a short reference to topological relationships is given, and a refinement is presented which groups the relationships into two clusters. The clusters are the basis for a distance function later.

In principle, it depends on representation of the space which (families of) topological relationships can be distinguished. In vector models a point set theoretic approach allows to define the following sets of a region X with the usual topology of the $\mathbb{R}^2$: the interior of X, the boundary of X, ∂X, and the exterior of X, $\neg X$. Refering to these sets, topological relationships between two regions can be defined by their nine intersection sets (Egenhofer and Franzosa, 1991). In raster models only X and $\neg X$ are distinguishable; but mapping a raster to the cited hybrid raster preserves full topology of $\mathbb{R}^2$ (Winter, 1995). A second factor influencing the number of relationships is the complexity of the regions (Egenhofer et al., 1994). – Relationships can be ordered in a conceptual neighborhood graph. For details we refer to the cited literature.

With regard to the vector and the hybrid raster representation, the distinguished relationships are: *disjunct*, *touch*, *overlap*, *coverage*, *containment*, and *equal*. – In (Winter, 1996) a refinement was introduced which is usefull also here: the relation *overlap* can be splitted into a *strong overlap* and a *weak overlap*, by a threshold of 50 % in the weight ϑ of overlap:

$$\vartheta = \frac{\|A \cap B\|}{\min(A, B)} \tag{1}$$

This refinement was used to partition the conceptual neighborhood graph into two relation clusters T_1 and T_2, which are connected only by the crossing from weak to strong overlap:

$$\begin{array}{rcl} T_1 & = & \{\textit{disjunct}, \textit{touch}, \textit{weak overlap}\} \\ T_2 & = & \{\textit{strong overlap}, \textit{coverage}, \textit{containment}, \textit{equal}\} \end{array} \tag{2}$$

With the assumption of small errors and uncertainties against the size of the regions, the assignment of a relation between two regions into a cluster is safe, having in mind that weak and strong overlap refer to the same topological relationship *overlap*.

4.3 A Detailed Distance Function Between Regions

In this section a distance function is described in detail, which characterizes *metric differences* between two similarly located regions[1] or unions of regions. The central idea is to observe distances of the two boundaries locally. A reference axis is required locating such distances uniquely. In the following we introduce a zonal skeleton for that reference axis, give a definition of a *local distance*, derive the *distance function* and its convolution, the *distance histogram*.

Consider Figure 5. On the left, two regions overlap considerably (relation $\in T_2$). In such cases let us concentrate on three intersection sets:

$$R = A \cap \neg B, \quad S = B \cap \neg A, \quad T = \partial A \cap \partial B \tag{3}$$

Through these (always connected) sets a medial axis is calculated (Serra, 1982), which we call a zonal skeleton. For an algorithm it is usefull to introduce here the complementary sets P and Q:

$$P = A \cap B, \quad Q = \neg A \cap \neg B \tag{4}$$

The zonal skeleton is the collection of center points of all circles that touch P as well as Q without intersecting P or Q. Then the *local distance* at any point of the skeleton can be defined by the (signed) diameter of its defining circle. The local distance is defined to be positive in R and negative in S (and 0 in T); it is *not* a metric. It is realized in the hybrid raster representation by a distance transform on $P \cup Q$, on the nodes of the hybrid raster. The

[1] A *non-metric* difference measure is presented in (Winter, 1998b).

distance function between A and B is the sequence of local distances along the skeleton (Fig. 5, right). If only the shape of the graph of the function interests, its origin ($x = 0$) is arbitrary. Otherwise one could define a start point at the skeleton point nearest to the origin of the reference system, or, in a raster, at the first occurence in search order.

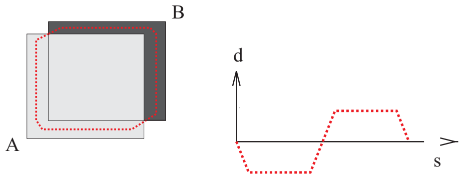

Figure 5: Two strong overlapping regions $r_{p1} = A$ and $r_{q2} = B$ (of square outline each) and the zonal skeleton (*left*), and the graph of local distances along the skeleton (*right*).

The described method works correct only for regions with a relationship in T_2. In other cases, i.e. for relationships in T_1, the zonal skeleton has to be defined through other intersection sets, but looks very similar; for details see (Winter, 1996). For complex regions, i.e. regions with holes, a hierarchic approach could be applied.

Several characteristics of differences between regions refer to the *histogram* of the distance function. The histogram is the density function of the distances. Let us assume a discrete distance function, as derived in the hybrid raster representation; for a continuous distance function it is necessary to resample the function in discrete steps. Then the histogram characterizes the frequency of local distances d along the skeleton:

$$h(d) = \|\{x \mid d(x) = d\}\| \tag{5}$$

The histogram is basis for other characteristics, like the distribution function, the mean, and moments of higher order.

5 CHARACTERIZING DIFFERENCES BETWEEN TWO REGIONS

In this section differences between two regions are investigated in two directions: topological differences are described qualitatively, and geometrical differences are described quantitatively.

5.1 Checking Topology

Basis for a comparison of the structure of matched, possibly complex regions are their region adjacency graphs (Section 4.1). Then a check of two region topologies means a comparison of two region adjacency graphs. Results are differences in the number of graph nodes and edges – the inner structure, or the level of partitioning –, and differences in node and edge attributes – the boundary structure, or level of detail (Fig. 2).

Problems in this procedure occur if some disjunct regions in one RAG match to the same region in the reference RAG. However, in this case the disjunct regions are to be connected, and the edges are to be labeled by *touch*.

5.2 Checking Geometry

Checking geometry is done by local distances in the hybrid raster representation (Section 4.3). See Figure 6: there are two tolerance levels introduced, a and b. Local distances inside of $[-a, +a]$ are considered as not significant, due to rounding, imprecision or noise, and local distances outside of $[-b, +b]$ are considered as indicating significantly different (parts of) regions. Intermediate values are treated as undecided. f_1 represents a distance function that exceeds partly the tolerance $+b$; that indicates that the associated regions differ in parts significantly, namely in R (Eq. 3): A outranges B in this area. In contrast, f_2 represents a distance function that is somewhere in between significant decisions. At last, f_3 represents a distance function that differs from 0 only by noise; the two regions can be classified as *equal* in geometry.

The same interpretation can be taken from the histogram of local distances (Fig. 7), because up to now only the extremal values of the distance functions are considered. Furthermore, additional information can be taken from the form of the distance curve and the histogram curve. In Fig. 6, f_1 exceeds *one time* the tolerance b: A outranges B in *one* part (this information is lost in the histogram). If the histogram shows two peaks symmetric to the y-axis and the distance function is smooth (like f_2), a shift exists between the two regions (Fig. 8). If the histogram shows two peaks, but one is centered, then support for the hypothesis exists that one building has additional parts (Fig. 9, left). If the histogram shows one centered peak but one-sided a certain amount of other distances, then probably one boundary node differs, and we classify to an outlier (Fig. 9, right). Other classifications can be made similarly.

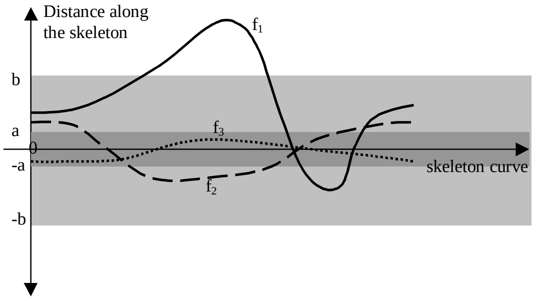

Figure 6: Three different classified distance functions. f_1: a pair of regions which most probably refer to different real world states (or are erroneous); f_2: a pair of regions with differences that are not automatically assessable; f_3: a pair of regions which is nearly equal.

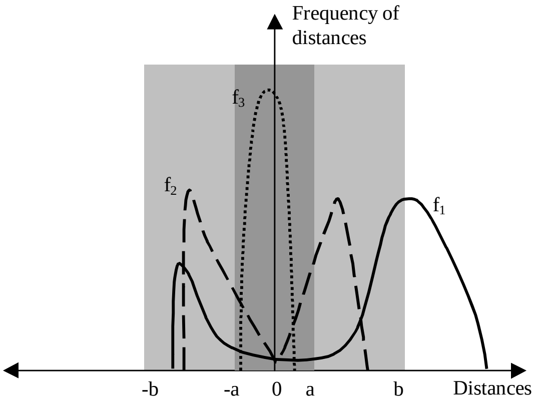

Figure 7: The histograms of the three different classified distance functions in Fig. 6.

The detailed geometric differences between associated regions can be generalized for the whole data set.

6 EMPIRICAL TEST

In this section the developed quality characteristics are applied to a comparison of two real data sets, where one is a reference for the other. Completeness of the second data set, differences in topological structure and in geometric location are investigated

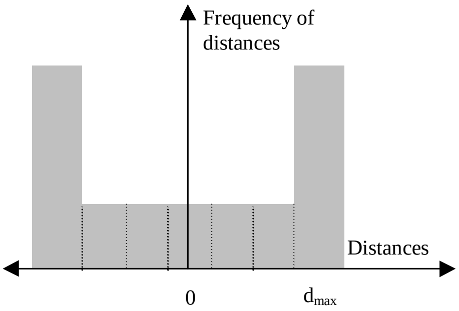

Figure 8: The histogram of a region shifted to the reference region.

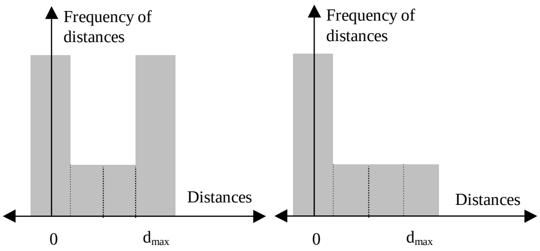

Figure 9: The histogram of a region covering the reference, i.e. with additional parts (*left*), and of a region with one boundary point different to (outside of) the reference (*right*).

and presented. It turns out that the quality descriptions are usefull to assess the second data set, and, by the way, its acquisition method.

6.1 Set-Up of the Test

Basis for the test are two data sets containing about 60 buildings of the same area. The first data set, $\mathcal{R}_1$, was produced from an image pair on an analytical plotter. We consider this data set as the reference because of the experience of the professional operator in image interpretation. The second data set, $\mathcal{R}_2$ is based on the same image pair, but was registered on a semi-automatic, model-based system for building extraction which has been developed at the Institute of Photogrammetry in Bonn (Gülch, 1997). Primitive volumetric models support image analysis, but could fit less to reality. For that reason we will describe the quality of this data set. It is supposed that there is no systematic error between the data sets (rotation, translation, scale), and both data sets are of a similar granularity.

The data sets contain simple, large and complex buildings from urban and industrial areas on the images. In detail, three areas are chosen and measured completely. The variation should minimize systematic influences of perspective, of building types or building complexity onto the quality descriptions.

The image pair is in a scale of 1:15.000. For the semi-automatic building extraction the images were scanned with $7\mu m$ pixel size; that corresponds to a ground resolution of 1 dm/pixel. With these numbers, let us assume a standard deviation of $\sigma = 1m$, i.e. 10 pixels.

The tolerances a and b (Sect. 5.2) can be chosen with regard to the standard deviation. The tolerance a, representing a threshold describing what is considered sufficiently near to be treated as equal, is set to σ, and the tolerance b, representing a threshold describing what is considered significantly different, is set to 3σ.

6.2 Results of the Test

Topology. The structural differences are taken from the region adjacency graphs. We have found out that there are systematic differences between a stereo-plotting data acquisition and a semi-automatic, model-based system. Typically, stereo-plotting yields less parts but more boundary points per part than the construction by primitives. Stereo-plotting supports topological data models; for that reason the only relationship between parts found in the *RAG* edge attributes is *touch*. Construction by primitives requires unification, and therefore also other topological relationships between parts are observed in the region sets from the model-based system. – Sometimes the partitioning level differs also by abstraction; Figure 11 gives an example of a simple building from stereo-plotting (below) that is composed by five primitive volumes in the second data set (above). The building is shown in Fig. 10.

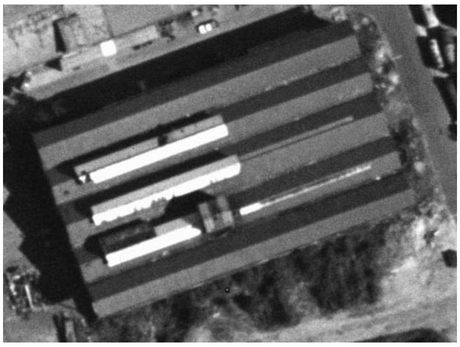

Figure 10: The building of Fig. 11 (©DeTeMobil 1998).

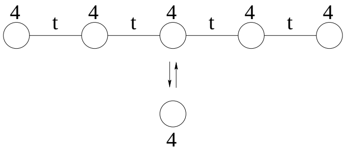

Figure 11: Two RAGs of a building: stereo-plotting (*below*) typically leads to less parts which are more detailed than model-based construction (*above*).

Geometry. For the geometry we have compared all the simple buildings and some complex buildings neglecting the partitioning. The differences are classified using the extremal values in the distance histograms and the tolerances a and b. The results for the data set are:

- *Correct factor* of 44% [30%-58%]: this factor is the percentage of matched regions with locational deviations inside of $[-a, a]$. The regions are considered as sufficiently identical, and the deviations are only due to numerical rounding effects, resolution higher than of practical interest, admitted inaccuracy, and so on.
- *Semi-error factor* of 37% [23%-52%]: this factor is the percentage of pairs of regions with locational deviations inside of $[-b, b]$. For an automatic system it is not possible to decide whether the compared regions represent the same real world object or not.
- *Error factor* of 19% [8%-32%]: this factor is the percentage of matched regions with locational deviations of a magnitude outside of $[-b, b]$. It is most likely that they do not refer to the same real world object, may be due to errors in data capture, may be due to changes in real world between two data captures (not in this test), or for any other reason.

The results in brackets indicate the 99% confidence region of the estimated percentages.

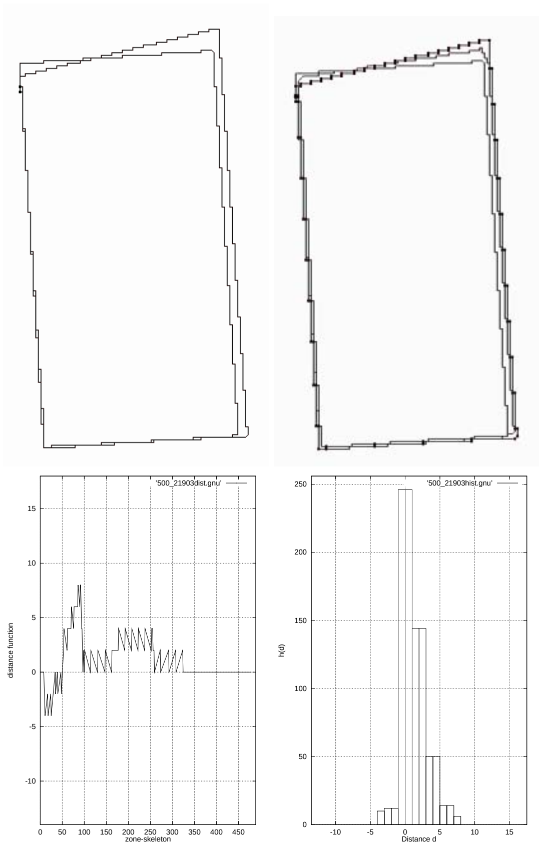

Figure 12: Outlines (*first*), skeleton (*second*), local distances (*third*), histogram (*last*) of a correct building ($|d| < 10$ pixels).

Buildings which fulfill the condition $\forall d : |d| < a$ have no (significant) geometric differences (Fig. 12). Other buildings can be classified by the type of the error, analyzing the form of the distance function and of the histogram.

If there is a secondary peak in the histogram then we have a difference in the location of at least a boundary point. Where the secondary peak is connected with the main peak by a constant density of significant length, the buildings differ in parts (Fig. 13). The sign of the local distance indicates which building is contained by the other one. – Two or more secondary peaks in the histogram indicate several differences in parts of the buildings. It depends on the distance value of the peak whether the difference is significant (Fig. 14). – An oscillating distance function indicates a comparison of a high-resolution building with a generalized one, which also may be caused by a model that does not fit to complex reality (Fig. 15).

7 SUMMARY AND CONCLUSIONS

We presented a systematic classification of differences between independently acquired regions, and we proposed methods for identification and quantification of those differences. Following the paradigm in spatial data base queries we investigate topology before analysing geometry in order to speed up, and also to clarify the analysis. Topological differences firstly are differences in the adjacency graph of the given regions. The geometric structure of the graph, e. g., the number of polygon points of the region boundary, is analysed next. This analysis, as well as the geometrical analysis of the boundaries, may refer to single regions as well as to corresponding sets of regions. The subsequent geometrical analysis uses the zone skeleton and its histogram.

We tested the procedure on 60 buildings acquired with two different methods and found it to be suitable for specifying the correctness or the differences of the two data sets. Statements about the completeness of the acquired data sets can be made additionally, after the topological and geometrical differences are identified.

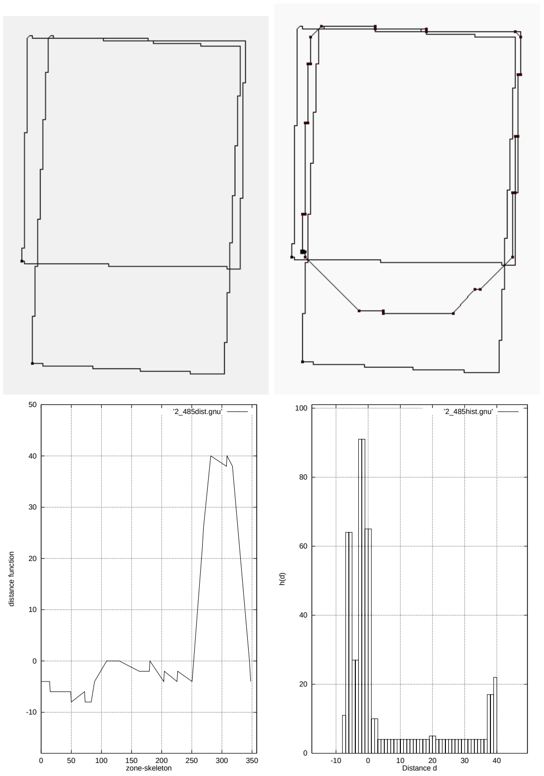

Figure 13: A significant difference in *abstraction*.

The geometric analysis using the zone skeleton has been automated, whereas the matching of the two region adjacency graphs stills waits for implementation. The final goal is to characterize the differences of two data sets as completely as possible, in order to identify typical failures in the acquisition procedures, and to reliably evaluate the differences with respect to given specifications. This may lead to clearer specifications of data acquisition procedures and increase the fidelity of automatic evaluation procedures.

Acknowledgement

We gratefully acknowledge the right of use of data sets from DeTe-Mobil, Bonn.

REFERENCES

Aalders, H. J. G. L., 1996. Quality metrics for GIS. In: Advances in GIS Research, Taylor & Francis, Delft, The Netherlands, pp. 5B.1–5B.10.

Baarda, W., 1967. Statistical Concepts in Geodesy. Vol. 2, Netherlands Geodetic Commission, Delft.

Bruns, H. T. and Egenhofer, M. J., 1996. Similarity of spatial scenes. In: M.-J. Kraak and M. Molenaar (eds), Advances in GIS Research, Taylor & Francis, Delft, pp. 173–184.

Burrough, P. A. and Frank, A. U. (eds), 1996. Geographic Objects with Indeterminate Boundaries. ESF-GISDATA, Vol. 2, Taylor & Francis.

Egenhofer, M. J. and Franzosa, R. D., 1991. Point-set topological spatial relations. International Journal of Geographical Information Systems 5(2), pp. 161–174.

D. Fritsch, M. Englich & M. Sester, eds, 'IAPRS', Vol. 32/4, ISPRS Commission IV Symposium on GIS - Between Visions and Applications, Stuttgart, Germany.

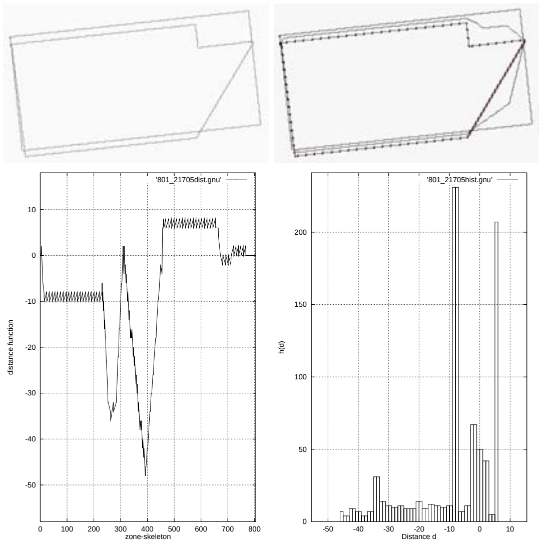

Figure 14: A significant difference by *generalization*.

Egenhofer, M. J., Clementini, E. and di Felice, P., 1994. Topological relations between regions with holes. International Journal of Geographical Information Systems 8(2), pp. 129–142.

Egenhofer, M. J., Flewelling, D. M. and Goyal, R. K., 1997. Assessment of scene similarity. Technical report, University of Maine, Department of Spatial Information Science and Engineering.

Fuchs, C., Lang, F., Förstner, W. and Vision, C., 1994. On the noise and scale behaviour of relational descriptions. In: H. Ebner, C. Heipke and K. Eder (eds), Proc. of ISPRS Comm. III Symposium *Spatial Information from Digital Photogrammetry*, SPIE, München, pp. 257–264.

Glemser, M., 1993. Untersuchungen zur objektbezogenen geometrischen Genauigkeit. Salzburger Geographische Materialien 20 pp. 97–108.

Goodchild, M. F. and Proctor, J., 1997. Scale in a digital geographic world. Geographical & Environmental Modelling 1(1), pp. 5–23.

Gülch, E., 1997. Application of semi-automatic building acquisition. In: A. Grün (ed.), Automatic Extraction of Man-Made Objects from Aerial and Space Images, Birkhäuser, Basel.

Guptill, S. C. and Morrison, J. L. (eds), 1995. Elements of Spatial Data Quality. Elsevier Science.

Haala, N., 1994. Detection of buildings by fusion of range and image data. In: ISPRS Comm. III Symposium on Spatial Information from Digital Photogrammetry and Computer Vision,, SPIE, pp. 341–346.

Harvey, F., Vauglin, F. and Ali, A. B. H., 1998. Geometric matching of areas. In: T. Poiker (ed.), Accepted Paper for Spatial Data Handling, Vancouver.

Joos, G., 1996. Assessing the quality of geodata by testing consistency with respect to the conceptual data schema. In: M. Craglia and H. Onsrud (eds), ESF-GISDATA and NSF-NCGIA Second Summer Institute in Geographic Information, Taylor & Francis, London, in press.

Knorr, E. M., Ng, R. T. and Shilvock, D. L., 1997. Finding boundary shape matching relationships in spatial data. In: M. Scholl and A. Voisard (eds), Advances in Spatial Databases (SSD '97), Vol. LNCS 1262, Springer, Berlin, pp. 29–46.

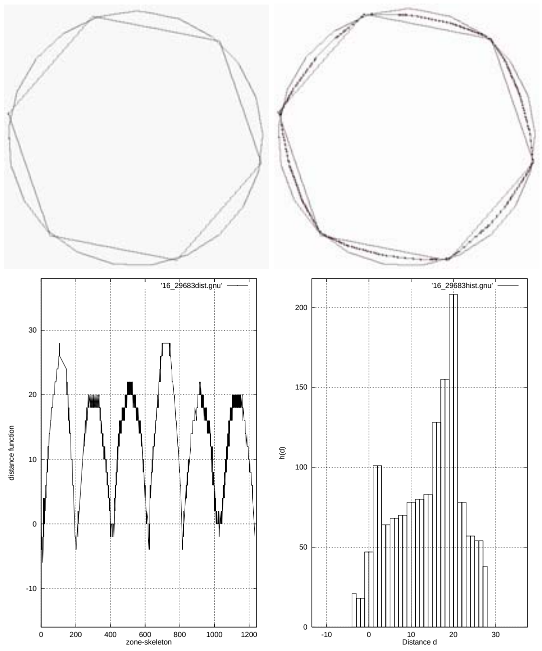

Figure 15: Limitations of a *primitive model* approach.

Kovalevsky, V. A., 1989. Finite topology as applied to image analysis. Computer Vision, Graphics, and Image Processing 46, pp. 141–161.

Peuquet, D., 1984. A conceptual framework and comparison of spatial data models. Cartographica pp. 66–113.

Serra, J. (ed.), 1982. Image Analysis and Mathematical Morphology. Vol. 1, Academic Press.

Stevens, S., 1946. On the theory of scales of measurement. Science 103(2684), pp. 677–680.

Timpf, S., Raubal, M. and Kuhn, W., 1996. Experiences with metadata. In: M.-J. Kraak and M. Molenaar (eds), Advances in GIS Research, Taylor & Francis, Delft, The Netherlands, pp. 12B.31 – 12B.43.

Tryfona, N. and Egenhofer, M. J., 1997. Consistency among parts and aggregates: a computational model. Transactions in GIS 1(3), pp. 189–206.

Tversky, A., 1977. Features of similarity. Psychological Review 84(4), pp. 327–352.

Winter, S., 1995. Topological relations between discrete regions. In: M. J. Egenhofer and J. R. Herring (eds), Advances in Spatial Databases, Lecture Notes in Computer Science, Vol. 951, Springer, pp. 310–327.

Winter, S., 1996. Distances for uncertain topological relations. In: M. Craglia and H. Onsrud (eds), ESF-GISDATA and NSF-NCGIA Second Summer Institute in Geographic Information, Taylor & Francis, London, in press.

Winter, S., 1998a. Bridging vector and raster representation in GIS. Internal report, Department of Geoinformation, TU Vienna.

Winter, S., 1998b. Location-based similarity measures for regions. In: ISPRS Commission IV Symposium, Stuttgart, Germany.