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

LOCATION-BASED SIMILARITY MEASURES OF REGIONS

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: similarity, matching, topological relations, spatial data quality, aggregation, generalization, assessment, change detection.

ABSTRACT

This paper contributes a systematic investigation of location-based similarity measures between discrete regions. It is shown that the number of measures is finite, and a complete list is presented and discussed. In literature one finds one or the other of these measures in use. But everywhere a practical approach leads to the measures, and no reference to alternatives is given. Here it is shown if – and to what extent – alternatives exist.
Similarity is a concept widely used; refering to space it is a basis for handling positional uncertainty or imprecision, for matching spatial entities, for merging spatial data sets, for change detection, for generalization, and so on. Similarity needs a measure to be quantifyable; a measure is the basis for any decision.
In this paper the focus is on similarity of a pair of discrete spatial objects, where only their location is considered. Location refers to the interrelation between position, shape and size of the objects. – Excluded from comparison are thematic attributes, relations in scenes, and matching. Furthermore, stochastic signals (as e.g. image regions) or fuzzy regions are not considered, but not excluded.
With the only precondition that a measure should be symmetric, normed, and free of dimension, area ratios are built and investigated. The list of ratios is complete, and only some of the ratios fulfill the precondition. These ratios are useful candidates for similarity measures, and their behaviour and semantical interpretations are discussed. Different measures refer to different reference systems; here, the different measures characterize different properties or interrelations between the position, the shape, and the size of two regions. None of the measures is a measure of overall similarity. Consequently, at least two of the listed measures are necessary. The two measures have to be complementary in the way that both together characterize common *and* distinct features of the regions.

1 INTRODUCTION

1.1 Motivation

Similarity is a concept widely used; referring to space and Geographic Information Systems (GIS) it is the basis for handling positional uncertainty or imprecision, for matching spatial entities, for merging spatial data sets, for change detection, for generalization, and so on. Since similarity is the basis for any decision in this context, it needs a measure to be quantifyable.

On the other hand, similarity represents an undecidable problem, which has been discussed in philosophy for two thousand years in the categorization controversy, where similarity is the central notion for abstraction (Flasch, 1986). The basic question is to find a common reference frame for measuring similarity: there are so many aspects of physical, linguistic or semantic similarity, that a statement *'A is similar to B'* contains no information as long as the aspects are not specified. For this reason the paper starts with a thorough clarification of *location* and *location-based similarity*.

Spatial entities in data bases, here assumed as regions, are models of real world objects. The comparison of the location of two regions from different data sets is based on the hypothesis that both are modeling the same object. The grade of similarity allows an assessment of that hypothesis. Conceptualizing the real world, a context-dependent level of detail, a dynamic world with changes, and random errors in data capture cause that models are most likely not identical. The differences can only partly be described stochastically, so methods like hypothesis testing are not helpfull in this situation. In contrast, similarity is a concept with a continuous measure between identity and distinctness. Specific measures yield additional information about the *kind of* similarity. In the context of comparing spatial data sets similarity yields a usefull tool for decision making.

1.2 Focus of the Paper

This paper contributes a systematic investigation of similarity measures between two discrete regions from different data sets (Fig. 1). Focus is on location, and it is assumed that the data sets represent some common space, so that location may correspond. Location refers to the interrelation between position, shape and size of the objects (Fig. 2). – Excluded from comparison are thematic attributes, relations in scenes, and matching. Furthermore, stochastic regions (as in image processing) or fuzzy regions are not considered, but not excluded.

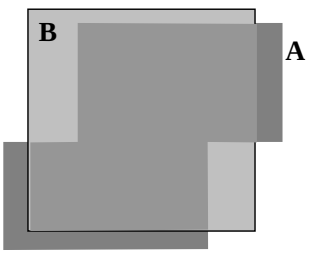

Figure 1: Given two regions A and B from two independent data sets: to what extent are they similar?

It is shown that the number of possible location-based similarity measures is finite (at least an elementary set), and a complete list is presented and discussed. In literature one finds one or the other of these measures in use. But usually a practical approach leads to these measures, and no reference to alternatives is given. Here it is shown if – and to what extent – alternatives exist.

With the only precondition that a measure should be symmetric, normalized, and free of dimension, area ratios are built and investigated. The list of ratios is complete, and only some of the ratios fulfill the precondition. These ratios are useful candidates for

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

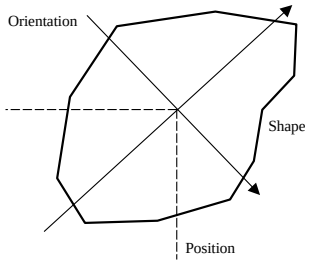

Figure 2: Location refers to the interrelation between position, shape and size.

similarity measures, and their behaviour and semantical interpretations are discussed. Different measures refer to different reference systems; here, the different measures characterize different properties or interrelations between the position, the shape, and the size of two regions. None of the measures is a measure of overall similarity. Consequently, at least two of the listed measures are necessary. The two measures have to be complementary in the way that both together characterize common *and* distinct features of the regions.

1.3 Structure of the Paper

The paper starts investigating similarity as a concept (Section 2), and clarifies location as a reference frame for similarity of regions (Section 3). Location-based measures are introduced based on intersection sets (Section 4). The size of the intersection sets is normalized by setting into ratios. These ratios are investigated and discussed in Section 5. In the conclusion (Section 6) this approach is discussed in a broader context.

2 SIMILARITY, AND SIMILARITY MEASURES

In this chapter concepts of similarity are collected. The overview is used to define the topic of this paper as well as to broaden the narrow view of geometry on similarity.

A concept of similarity exists:

- in mathematics: a transformation $h : S \rightarrow T$ is a similarity if there is a constant number r such that $f(h(x), h(y)) = r\, f(x, y)$ for all $x, y \in S$ (Edgar, 1990). For example, two geometric figures are similar if the ratio of all related pairs of edges is constant. – This concept of similarity does not refer to the size of r and is not quantifyable.

  Another concept is congruence, but that does not allow gradation. But in topology similarity can be applied; (Bruns and Egenhofer, 1996) derive similarity measures of scenes by conceptual neighborhoods of relations.

- in statisitics: two signals are correlated (or linear dependent) if their covariance is different from 0. Similarity of signals is described often by the cross correlation coefficient, a symmetric, normalized correlation measure (e.g. in (Jähne, 1995)). The coefficient is based on the distance of the two signals. – This concept is more strict than the common sense concept, which does not require dependency.

- in common sense: proximity, or closeness to equality (which *per se* does not refer to coincidence in location). It depends whether one is considering the common features (*similarity*) or the distinct features (*dissimilarity*) which basis for a similarity measure is chosen. Often the *distance* from equality is accepted as a measure for similarity, as in statistics. – This concept is open for continuous or discrete gradation or order (*'A is more similar to B than to C'*).

- in psychology / cognitive science: similar things belong to categories which are characterized on the basis of shared properties (with some additional aspects) (Lakoff, 1987). Categorization tends to count the common features. Tversky postulates that similarity of objects increases with the number of common features, and decreases with the number of distinct features (Tversky, 1977). This concept exceeds also the traditional concepts based on distances because a distance controls distinct features only. – Tversky's rule influences the argumentation in this paper strongly.

  Similarity of visually perceived spatial objects is a topic of Gestalt-theory (Metzger, 1936).

- in neuropsychology: neurons are trained by repetition of the same (or similar) signals, and response to a stimulus is proportional to the grade of similarity of an input signal to the trained pattern. – This concept of similarity is compatible with counting common features; it produces a continuous measure of similarity.

- in applied sciences (engineering, operation research, etc.): two entities, or two situations are similar with a certain degree determined by the costs of mapping one onto the other. Costs are related to distances: only mapping of distinct features raise costs. As Vosselman pointed out, sometimes it is more usefull to determine benefits instead of costs (Vosselman, 1992); mapping common features would increase the benefits. Nevertheless, the considered features are not necessarily numeric (Stevens, 1946), which requires additional efforts to determine the measure function.

It becomes evident that the concepts vary, and no overall concept emerges. There are so many aspects of physical, linguistic or semantic similarity, that a statement *'A is similar to B'* contains no information as long as the aspects are not specified which are referred to. Collecting similarity measures makes sense only for general classes of aspects, as done here for *location*. As we will see, the derived measures will describe even different aspects of location (Sect. 5.2).

Another conclusion from the derived measures is evidence that *two* measures are required to describe similarity completely. Similarity is regards the number of common properties – here: coincident location –, as well as the number of distinct properties. The number of common properties characterizes *similarity* (in a narrower sense), contrasted with the number of distinct properties that characterizes *dissimilarity*.

Notational remark: The notion *similarity* is used in this paper as a general concept as well as a concept regarding common properties only.

3 SIMILARITY OF REGIONS

In this section the discussion of similarity is focused on regions. The concept of location is presented, and sources for deviations in location are discussed. At the end we have good reasons to choose elements for similarity measures.

3.1 Location of Regions

In this paper no assumptions are needed about history, context, or quality of the two considered data sets. They are presumed as referring to the same part of the world, and no explanations are given or derived for locational deviations between the objects of the two data sets (as e.g. different context, different level of detail, different observers, ...). Similarity is only considered in relation to location here.

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

Location of regions can be characterized by a large number of parameters describing their shape, their orientation, and their position individually (cf. Fig. 2):

- Measures of shape, like the Euler number, or the size of the region;
- measures of orientation, like the angle between reference and inertia axis;
- measures of position. In two-dimensional space that is a pair of coordinates in a suited reference system. The reference system in GIS is often $\mathbb{R}^2$ (but non-euclidic (spheric) or discrete ones are also possible).

The concept of location is related to coverage of space; the set of points or atoms occupied in the plane by the considered regions are compared, not their attributes (like shape, orientation or position). However, location includes shape, orientation and position indirectly: a change in shape means a change in location at some points in the plane, and so on.

Further two regions from different data sets are compared neglecting (problems by) representational differences. The following argumentation is independent from representational issues.

3.2 Sources for Different Locations

Consider the observation process (Fig. 3). The continuous real world is mapped by perception and abstraction to a conceptual model, and that model is mapped by data capture to a data set. Comparing objects from two data sets requires the inversion of some paths in the two branches in that hierarchy.

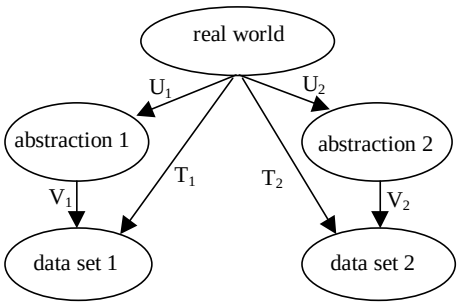

Figure 3: Mapping two data sets requires to go back to common ground, here: to the real world.

The mapping process from the real world to a data set is called T_k in Fig. 3, with $k \in \{1, 2\}$. The mapping is a process of two distinct steps, an abstraction of continuous real world to discrete concepts (U_k) and a measurement of conceptualized objects (V_k).

$$T_k = V_k.U_k$$

Looking for a mapping function between the data sets S_k requires inversion of one of those derivation processes, e.g. by

$$S_2 = T_2.T_1^{-1}(S_1)$$

This task is ill-posed by:

1. irreversible mental processes, especially conceptualization and simplification (abstraction in U_k),
2. dimensional aspects (data sets are assumed to be two-dimensional, the real world is (at least) three dimensional) (projection in U_k),
3. observation errors (in V_k).

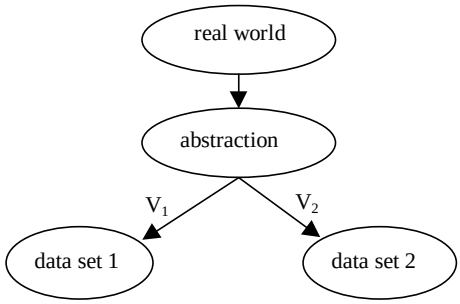

Figure 4: If two data sets are based on a common conceptualization of the world, all differences between the data sets are due to observation errors. Only then the derivation process should be reversible.

The second problem can be constrained by parallel projection. The last problem can be modeled stochastically, which allows formulating inversion by an estimation problem. The first problem, however, is irreversible.

Figure 4 shows a reversible case because here the only reason for differences between two data sets are observation errors. Possible differences can be modeled stochastically.

$$S_2 = V_2.U.U^{-1}.V_1^{-1}(S_1) = V_2.V_1^{-1}(S_1)$$

A practical example for this figure may be an operator who uses a stereo pair of images to derive two building data sets in sequence by the same capturing method. But if he uses two different methods, let us say, stereo-plotting and construction by primitive volumes (CSG), then non-stochastic constraints by object models influence the results.

Usually we have to deal with the general case (Fig. 3). With the lack of a mathematical transformation between S_1 and S_2, matching of data sets is referred back to empirical methods, e.g. on decisions based on similarity measures. When applying empirical methods, however, it is difficult to give theoretical reasons. The measures investigated here are based on local coincidences and differences, which are perhaps the most general similarity of (models of) spatial objects. To remain general we neglect all other possible measures here, and any possibility of analysing the types of differences (shift, rotation, affinity, ...).

It is clarifying to distinguish similarity of regions at different levels. Similarity is a binary relation; the two arguments can:

- refer to different objects in the real world. In this case similarity concerns shape only, and the difference in position or orientation will not be considered.
- refer to the same object, but to different concepts of abstraction. In this case similarity concerns the two contexts, and location-based measures can be used to specify one type of describing indicators.
- refer to the same object class and abstraction level, but to different data sets. In this case similarity in location concerns identity, or at least part-of relations. Similarity will be used to match regions, to detect differences in data sets, e.g., changes, and so on.

In special circumstances the third case can be treated as an estimation problem of a shift between two correlated (spatial) signals. That is common practice in image matching (Ackermann, 1984). But then the distinct properties between the two signals (regions) break the model und must therefore be small. In this paper it is avoided to make any restriction about the shape or a correlation between the two compared regions. For that reason matching techniques are not considered further.

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

We classify the reasons for different locations of two matching regions into:

1. Uncertainty
   - *in abstraction:* when observers differ significantly in their spatial concepts (cf. (Burrough and Frank, 1996)), or in the level of detail for a concept.
   - *in measurements:* clearly (and uniquely) defined concepts measured several times will differ in their description (data sets), due to measurement methods, instruments, observers, and random influences.
2. Errors
   - *gross error.* Gross errors prohibit correct matching of two regions from two data sets. Gross errors occur in shape: e.g., an 'outlier' in one boundary polygon, and gross errors occur in position or orientation: e.g., if the size of shift or rotation is of a magnitude that dissimilarity predominates. Then the similarity measures presented in this paper will be low; detection and elimination is possible only in global optimization with robust methods if a majority of matches is gross-error-free.
   - *systematic error.* Systematic errors of location can be caused by the observation method, e.g., shadows in automatic object extraction (Weidner and Förstner, 1995). They cause a dissimilarity trend in the comparison of all pairs of objects between two data sets (shift, rotation, affinity, scale, and so on). But similarity measures fall short to derive more detailed information, because they do not allow to distinguish shape, position, and orientation, and they are not signed. For another approach of comparing regions see (Ragia and Winter, 1998) in this volume.
   - *random error.* Random errors break equality of regions to any kind of similarity in any of the dimensions shape, position, and rotation. Random errors are typically small, so that similarity measures will be high.
3. Temporal aspects

   GIS databases contain immobiles – movable regions are usually not stored. But also the world of immobiles is not static, and databases represent regions with reference to a point (or interval) in time (Snodgrass, 1992). Comparing regions from two datasets requires consideration of database time:
   - *creation* or *deletion* of regions in one data set lead to mismatches (or no matches) with other data sets.
   - *change* of regions in a static space can be reduced to creation and deletion. Examples are creation or deletion of parts of a region (construction of an annex), or deletion of an individual region and creation of a new one (displacements; division of parcels). – Other changes in a data set refer not to the real world, but to reference frames; they base on recalculation of networks or transformations. They keep identities, but require a complete database revision.

     Changes regard all aspects of location.

   Temporal aspects give reason for dissimilarities at a descriptive level; the observable phenomena – similarity measures – look like gross or systematic errors, *significant* changes presumed.

Up to now, we discussed the location of regions, and reasons for differences between data sets, even if they refer to the same objects and abstraction level. We saw that matching of data sets is an ill-posed problem which requires heuristics, i.e. empirical approaches.

3.3 Approach for Location-Based Measures

Location of an object is understood as space covered by that object (Fig. 1). Measures based on location can count (or integrate) atomic elements of space; that are points in $\mathbb{R}^2$, and raster cells in $\mathbb{Z}^2$. Binary location-based similarity measures will count atoms covered by both objects, or atoms covered by one but not by the other object.

Location-based similarity is a concept related to the topological relationship *equal*, in a sense of *more or less*. The strong mathematical formulation of topological relationships (e.g. by emptyness / non-emptyness of intersection sets, (Egenhofer and Herring, 1990)) can be softened to graded or fuzzy or uncertain topological relationships (Wazinski, 1993, Winter, 1996). For that aim the sizes of intersection sets are combined to ratios.

This paper states that the number of such measures (at least the elementary set) are finite. The next chapter collects location-based measures, and the consecutive chapter investigates the ratios of those measures.

4 LOCATION-BASED MEASURES

In this chapter we derive the location-based measures, with special attention to be complete. They will be based on the sizes of (intersection) sets, with a strong interrelation to weighted topological relationships.

4.1 Partition of the Plane

Studying relative location refers at the most general level to topology. In this section topological relationships are characterized by two-dimensional intersection sets. Later the qualitative relationships will be specified by the size of sets.

It is generally assumed in the following that all treated areal objects are existing and not empty. Location of areal objects is represented in GIS usually as bounded parts of the plane (*vector model*), or as sets of (regular) atoms (*raster model*, *tesselation*). Here an approach by a location function is preferred:

$$f(x,y) = \begin{cases} 0 & \text{if } (x,y) \notin A \\ 1 & \text{if } (x,y) \in A \end{cases} \quad (1)$$

with $(x,y) \in \mathbb{R}^2$ (vector model) or $(x,y) \in \mathbb{Z}^2$ (raster model), respectively. That keeps us independent from representational issues. Without loss of generality, in the following it is referred to $\mathbb{R}^2$ only. But the formulas can be applied to $\mathbb{Z}^2$, too, by replacing integrals by sums.

Changing from an object view to a location view (Winter, 1998) is the first step of solving the geometric matching problem. If one is not interested in all other aspects of comparing, but only in position, then it is sufficient to distinguish space between 'region' (foreground) and 'no region' (background).

The location function (Eq. 1) distinguishes two sets, the interior ($f(x,y) = 1$) and the exterior ($f(x,y) = 0$) of an area A. The function needs no concept of neighborhood. Therefore, open and closed sets cannot be distinguished in the functional representation. The inverse of function f, f^{-1}, yields the complement of A, i.e. $\neg A$. For two areas, A and B, a set of in total four (intersection) sets can be derived. Consider Figure 5. Region A from a data set **A** and region B from a data set **B** have an arbitrary position relative to each other (in the figure they are overlapping, and the rectangle A is left of the rectangle B). Their intersection sets form a partition of the planar space with at most four sets:

$$\begin{array}{rcl} A & \cap & B \\ \neg A & \cap & B \\ A & \cap & \neg B \\ \neg A & \cap & \neg B \end{array} \quad (2)$$

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

All other sets are unions of those intersection sets. For example:

$$\begin{array}{lcl} A & = & (A \cap B) \cup (A \cap \neg B) \\ B & = & (A \cap B) \cup (\neg A \cap B) \\ A \cup B & = & (A \cap B) \cup (\neg A \cap B) \cup A \cap \neg B) \end{array}$$

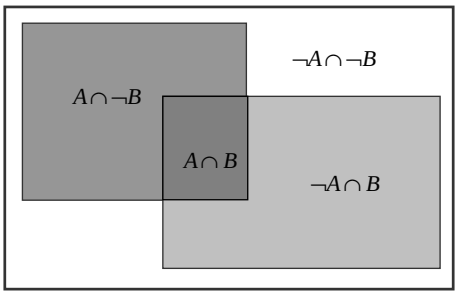

Figure 5: Sets A and B given as rectangles; their intersection sets form a partition of the plane. Background is assumed to be unlimited.

The main interest is in the size of the sets. An elementary operation *sizeof* is introduced here, shortly written in mathematical notation by $|\cdot|$. The application of this operation on sets in $\mathbb{R}^2$ shall yield their size, independent from the representation of the sets. We define:

$$\begin{array}{lcll} \mu_{11} & = & |A \cap B| & = & \iint_{x,y} f(x,y)\, g(x,y)\, dx\, dy \\ \mu_{12} & = & |A \cap \neg B| & = & \iint_{x,y} f(x,y)\, g^{-1}(x,y)\, dx\, dy \\ \mu_{21} & = & |\neg A \cap B| & = & \iint_{x,y} f^{-1}(x,y)\, g(x,y)\, dx\, dy \\ \mu_{22} & = & |\neg A \cap \neg B| & = & \iint_{x,y} f^{-1}(x,y)\, g^{-1}(x,y)\, dx\, dy \end{array} \quad (3)$$

Once the sizes are known, (families of) topological relationships can be distinguished. To become qualitative, the four size measures of Eq. 3 are classified and grouped into tuples:

- The sizes of intersection sets, μ (Eq. 3), can be mapped to a binary measure m with values 0 = *empty* and 1 = *non-empty*:

$$\mu \to m = \begin{cases} 1 & \text{if } \mu \neq 0 \\ 0 & \text{if } \mu = 0 \end{cases} \quad (4)$$

- The size μ_{22} (Eq. 3) is – for finite A and B – never empty. It contributes no qualitative information. With unlimited functions f and g, its size is always ∞, so it contributes even no quantitative information. m is 1, constantly.

  Then a situation between regions A and B can be described qualitatively by combinations of the binary measure m, where it is sufficient to set up triples $\{m_{11}, m_{12}, m_{21}\}$. That yields $2^3 = 8$ theoretically possible combinations.

- No pair of (m_{11}, m_{12}) and (m_{11}, m_{21}) can be (empty, empty). That follows from the presumption that neither A nor B is empty. With $A = \mu_{11} \cup \mu_{12}$ and $B = \mu 11 \cup \mu_{21}$ at least one of the intersection sets in each pair must be not empty. That excludes three of the eight triples: $\{0,0,0\}$, $\{0,1,0\}$, $\{0,0,1\}$.

The remaining five triples correspond to the following separable topological relationships:

1. $\{0,1,1\}$ (disjunct/touching): A and B have no part in common;
2. $\{1,1,1\}$ (overlap): A and B have parts in common and parts not in common;
3. $\{1,0,0\}$ (equal): all parts of A are parts of B and vice versa;
4. $\{1,1,0\}$ (contains/covers): all parts of B are part of A, and A has additional parts;
5. $\{1,0,1\}$ (containedBy/coveredBy): all parts of A are part of B, and B has additional parts.

Using four intersection sets looks similar to the work of (Egenhofer and Franzosa, 1991) who determined the topological relationship between A and B qualitatively. But they investigated the intersection sets of interiors and *boundaries* with the result that they can separate eight (families of) topological relationships for simple areas.

While Egenhofer and Franzosa were restricted to simple areas, the classification here works also for complex areas (multiply-connected, or multiple components). The used intersection sets are of the dimension of the considered space. For that reason the model is indepent from representation (vector or raster). With this subset of the Egenhofer relations Fig. 6 is a kind of generalization of the conceptual neighborhood graph (Egenhofer and Al-Taha, 1992).

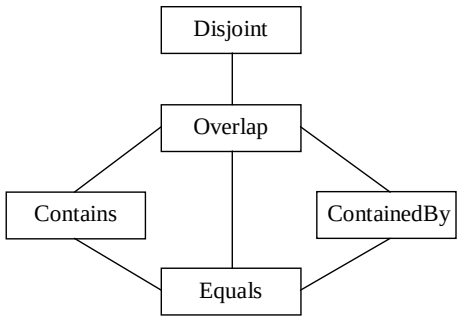

Figure 6: Topological relationships representable by the two-dimensional intersection sets of Eq. 3, related by conceptual neighborhood.

We are now able to describe an area by a function, and to determine a topological relationship between two areas qualitatively. The next step is to become quantitative.

4.2 Size Measures

In this section the size measures of Eq. 3 are investigated numerically.

The domain of values for the size of an arbitrary set X in $\mathbb{R}^2$ is dom $(|X|) = [0, \infty]$. But regions A and B shall be limited to finite sets which may not be empty (Fig. 5). Then $0 < |A|, |B| < \infty$.

We derive the sizes of intersection sets:

$$\begin{array}{rcrcl} 0 & < & |A \cap B| & \leq & \min(|A|, |B|) \\ 0 & < & |\neg A \cap B| & \leq & |B| \\ 0 & < & |A \cap \neg B| & \leq & |A| \\ & & |\neg A \cap \neg B| & = & \infty \end{array} \quad (5)$$

All other sets are unions of intersection sets; with the property of partitions to be pairwise disjoint the sizes of unions can be written as sums:

$$\begin{array}{lcl} |A \cup B| & = & |A \cap B| + |\neg A \cap B| + |A \cap \neg B| \\ |\neg A \cup B| & = & |A \cap B| + |\neg A \cap B| + |\neg A \cap \neg B| \\ |A \cup \neg B| & = & |A \cap B| + |A \cap \neg B| + |\neg A \cap \neg B| \\ |\neg A \cup \neg B| & = & |A \cap \neg B| + |\neg A \cap B| + |\neg A \cap \neg B| \end{array} \quad (6)$$

In the following we use unions as short forms for the combination of intersection sets. The domains of the binary unions are easy to determine from the domains of the intersection sets. If one term of the sum in Eq. 6 is ∞, the domain is fixed to ∞, too. Only the first union is a finite set:

$$\begin{array}{rcrcl} \max(|A|, |B|) & \leq & |A \cup B| & \leq & |A| + |B| \\ & & |\neg A \cup B| & = & \infty \\ & & |A \cup \neg B| & = & \infty \\ & & |\neg A \cup \neg B| & = & \infty \end{array} \quad (7)$$

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

In this section we collected the sets and the domain of their sizes. The size measure is based on integration. It is shown that the elementary sets are complete. – In a next step we establish two criteria for similarity measures, and derive such measures from the size measures.

5 LOCATION-BASED SIMILARITY MEASURES

In this chapter we derive location-based *similarity* measures, with special attention on completeness. The size measures of Chapter 4 are used and coupled with three criteria for similarity measures – to be symmetric, normalized, and free of dimension. It will be possible to set up lists of such measures and to describe their properties.

5.1 Criteria for Similarity Measures

In this section three criteria are established to specify similarity measures. With these criteria it will be possible to derive such measures from the size measures.

The criteria are:

1. Symmetry

   From our assumptions the situation between A and B is symmetric; no region is preferred as e.g. a prototype or a model of the other. In such neutral situations a measure must be independent from the order of the considered regions A and B:

   $$\text{similar}\,(A,B) \stackrel{!}{=} \text{similar}\,(B,A) \tag{8}$$

2. Domain limitation

   It is usefull to have normalized measures. Only in this case two measures can be compared.

   $$0 \leq \text{similar}\,(A,B) \leq 1 \tag{9}$$

   For this reason suited *ratios* are introduced as similarity measures. (Suited ratios and their meaning of such ratios are discussed in Section 5.2.)

3. Free of dimension

   Similarity measures shall be free of dimension, because similarity is no physical concept or property. That is reached by ratios of measures with the same dimension.

First we consider symmetry in the size measures. With the partition into (at most) four intersection sets, in principle 16 combinations are possible. The numbers follow from the sequence of binomial coefficients $\binom{n}{k}$, with $n = 4$, the number of intersection sets, and $k \in \{0, ..., 4\}$, the number of combined elementary sets:

- $k = 0$: one 0-tuple, the empty set (excluded by presumption);
- $k = 1$: four 1-tuples, the elementary intersection sets;
- $k = 2$: six 2-tuples, binary unions of intersection sets;
- $k = 3$: four 3-tuples, triple unions of intersection sets;
- $k = 4$: one 4-tuple, the union of all four intersection sets, equal to $\mathbb{R}^2$.

The cases with $k = 0$ and $k = 4$ are meaningless in the context of similarity. From all other tuples only a few are symmetric (taking advantage from abbreviations by unions, cf. Eq. 6):

- $k = 1$: $A \cap B$, $\neg A \cap \neg B$
- $k = 2$: $(A \cap \neg B) \cup (\neg A \cap B)$, $(A \cap B) \cup (\neg A \cap \neg B)$
- $k = 3$: $A \cup B$, $\neg A \cup \neg B$

In the following it is sufficient to investigate this reduced set of sets, as the only possible symmetric sets. We pass over to their sizes, and investigate the ratios of size measures to find normalized similarity measures (Eq. 9).

For that purpose the domains of size values are used (Eqs. 5, 7). Then directly follows that three of the six measures cannot be normalized because they are fixed to ∞. The remaining three sizes are:

$$\begin{array}{l} |A \cap B| \\ |A \cap \neg B| + |\neg A \cap B| \\ |A \cup B| \end{array} \tag{10}$$

Looking at their value domains show the requirement of two additional size measures as upper bounds: $\min(|A|, |B|)$, and $|A| + |B|$. A third measure, $\max(|A|, |B|)$, is linear dependent with the other two, by $|A| + |B| = \min(|A|, |B|) + \max(|A|, |B|)$. Then it is sufficient to introduce $\min(|A|, |B|)$ and $\max(|A|, |B|)$. All three of these measures are independent from location, and for that reason they are not considered as candidates for location-based similarity measures. Besides, these measures are symmetric.

Remark. Another symmetric, location-invariant measure exists: $|A| \cdot |B|$. This measure has the value domain $0 < |A| \cdot |B| < \infty$. It would be usefull for setting up normalized ratios (in the next section). But it is of a higher dimension than the measures above, and is therefore excluded.

5.2 Composing the Similarity Measures

In this section the remaining size measures are normalized, and the resulting ratios are investigated in their meaning. The combinations of nominators and denominators will be complete for all *location-sensitive* measures in nominators.

With three location-sensitive size measures (Eq. 10) the list of nominators contains three elements. Measures for the denominator may never take the value 0. This argument excludes the measures $|A \cap B|$ and $|(\neg A \cap B) \cup (A \cap \neg B)|$ from the list of possible denominators. With six measures altogether, and two excluded measures, the list of denominators contains four elements. Therefore, a matrix of 3×4 ratios is to be investigated now (Tab. 1).

nominator / denominator	$\|A \cap B\|$	$\|\neg A \cap B\| + \|A \cap \neg B\|$	$\|A \cup B\|$
$\|A \cup B\|$	s_{11}	s_{12}	s_{13}
$\min(\|A\|, \|B\|)$	s_{21}	s_{22}	s_{23}
$\max(\|A\|, \|B\|)$	s_{31}	s_{32}	s_{33}
$\|A\| + \|B\|$	s_{41}	s_{42}	s_{43}

Table 1: Combination of all possible ratios of size measures.

In detail:

s_{11} Domain of values $[0, 1]$. 0 stands for totally disjoint regions ($A \cap B = \emptyset$), and 1 stands for identical regions ($A \cap B = A \cup B$). This ratio is a prototypical example of a location-based similarity measure, increasing with the grade of similarity.

s_{12} Domain of values $[0, 1]$. 0 occurs only if $A = B$, and 1 occurs if A and B are totally disjoint. With this behaviour the ratio complements s_{11}, which bases on the complementing nominators with regard to the denominator. One should call it a dissimilarity measure, decreasing with the grade of similarity.

s_{13} Domain of values $[1]$, trivially.

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

s_{21} Domain of values $[0, 1]$. 0 stands for totally disjoint regions, and 1 stands for complete coverage/containment or identity ($\subseteq$). The ratio does not recognize the proportion in size between A and B, and therefore it is not suited as a similarity measure. But this ratio could be used as a measure for the grade of (symmetric) overlap.

s_{22} Domain of values $[0, \infty)$. Again, 0 occurs only if $A = B$. But the denominator is not sufficient to normalize the nominator. That property excludes this ratio from the list of similarity measures. Additionaly, values different from 0 are difficult to interpret, because nominator and denominator are not correlated.

s_{23} Domain of values $[1, \infty)$. 1 occurs if $A = B$, and the ratio increases in all other cases. With not being normalized, this ratio is excluded from the list of similarity measures.

s_{31} Domain of values $[0, 1]$. 0 occurs if both regions are disjoint, and 1 occurs only if $A = B$, in contrast to s_{21}. With its sensitivity for proportions between A and B this ratio is a suited similarity measure.

s_{32} Domain of values $[0, 2]$. 0 occurs if $A = B$, and 2 occurs if A is disjoint from B and $|A| = |B|$. As long as one region is covered/contained in the other region, the value of the ratio is limited by an upper bound of 1. As long as both regions are disjoint, the value of the ratio is limited by a lower bound of 1. In any case of overlap no prediction can be made. – This ratio could be normalized by division by 2; then it represents a dissimilarity measure (decreasing with growing similarity).

s_{33} Domain of values $[1, 2]$. The value 1 stands for all cases of coverage/containment or identity. The value 2 occurs for disjoint regions, if $|A| = |B|$. – Neither the domain nor the behaviour recommends this ratio as a similarity measure.

s_{41} Domain of values $[0, \frac{1}{2}]$. 0 stands for disjoint regions, and $\frac{1}{2}$ stands for $A = B$. If we would normalize the ratio (by multiplication with 2), the result would be a mean size of A and B as denominator ($\frac{|A|+|B|}{2} = \frac{\min(|A|,|B|)+\max(|A|,|B|)}{2}$. With that the behavior of the (normalized) ratio s_{41} is in between of s_{31} and s_{21}. It yields no new information.

s_{42} Domain of values $[0, 1]$. 0 occurs if $A = B$, and 1 occurs if A and B are disjoint. Again, this is a mean ratio of s_{22} and s_{32}, but this one fulfills the conditions of a (dis-)similarity measure.

s_{43} Domain of values $[\frac{1}{2}, 1]$. The lower bound occurs if $A = B$. 1 occurs in all cases of disjoint regions, but is reached also in all other topologic relations, if $|A|$ and $|B|$ are different in the order of magnitude. This ratio represents an extraordinary dissimilarity measure.

In summary, from the possible ratios of size measures the following are similarity measures: $\{s_{11}, s_{31}, s_{41}\}$, and another list are dissimilarity measures: $\{s_{12}, s_{32}, s_{42}, s_{43}\}$. Both lists are complete with regard to the given criteria.

5.3 Combination of Similarity Measures

In this section combinations of similarity measures are investigated. Evidence is given that both lists above are needed, which is supported by some examples of recent applications (Harvey et al., 1998).

Tversky already postulates that similarity of A to B *"is expressed as a function h of three arguments: ... the features that are common in both A and B; ... the features that belong to A but not to B; ... the features that belong to B but not to A."* ((Tversky, 1977), p. 330). – With this argumentation in mind, our lists of similarity and dissimilarity become more transparent. All similarity measures are based on the nominator $|A \cap B|$, which represents the common features between A and B. All dissimilarity measures, with one exception, are based on the nominator $|\neg A \cap B| + |A \cap \neg B|$, which represents the distinctive features of A and B. The exception, s_{43}, treats topological relations combined with orders of magnitude, which mixes different kinds of features, metric and topologic ones.

These considerations lead to the expectation that in praxis one measure from each list is required to assess similarity completely.

Consider a recent example (Harvey et al., 1998). To evaluate a match of two regions they introduce two measures: an *inclusion function*, which is in fact identical to s_{21} and yields the grade of overlap instead of similarity (but nevertheless: the common features), and a *surface distance*, which is identical to s_{12} and measures dissimilarity (distinctive features). That confirms the hypothesis that two measures are needed. The interesting question remains whether other pairs of measures would have been also usefull. The authors do not discuss their choice.

Another example is mentioned in (Ragia and Winter, 1998). There the matching of two buildings from two data-sets has special requirements, with regard to the aggregation levels of the data-sets. part_of-relations are accepted as a match. Similarity is replaced by weighted topological relations, e.g. by s_{21} and s_{31}. With this choice distinctive features are not considered, only common features.

Similarity of regions is to be handled distinctly to similarity of lower dimensional entities. Recently, (Walter, 1997) matches lines and points. He works only with distance measures (costs), neglecting a weight for common features. That is justified for one-dimensional data-sets, because the probability that two lines coincide by chance is very small (the probability for two points is even zero).

Similarity of spatial relations cannot be treated by sizes of sets (the single exception are topological relations). For example, (Bruns and Egenhofer, 1996), (Egenhofer, 1997) are investigating spatial scenes. Though they involve metric refinements of topological relations (cf. Eq. 3), they need an additional concept of similarity for other spatial relations. They also work with distance measures, which they derive from conceptual neighborhood graphs.

6 SUMMARY, DISCUSSION AND CONCLUSION

This paper presents a systematic investigation of location-based similarity measures between discrete regions of different data sets. It is shown that only seven of such measures exist if only measures are considered which are symmetric, normalized, and free of dimension. The set of similarity measures can be classified into the measures counting common features of regions, and measures counting distinct features. A complete description of similarity requires one measure from both classes.

With measuring the sizes of intersection sets some similarity measures are related strongly to graded topological relationships. s_{11} represents a grade of *equals*, s_{21} represents a grade of *overlaps*, s_{12}, as the complement of s_{11}, represents a grade of *disjoint*. Gradations of containment cannot be found; but a concept of a graded containment may coincide with the grade of overlap, intuitively. Boundary based topological relationships are not treated here.

Tversky proposes a *contrast model* which expresses similarity between objects as a weighted difference of the measures of their common and distinct features (Tversky, 1977):

$$\begin{aligned} s(A,B) &= h(A \cap B, A - B, B - A) \\ &= \alpha h(A \cap B) - \beta h(A - B) - \gamma h(B - A) \end{aligned} \quad (11)$$

The advantage of that approach is to have only one measure for overall similarity. But on the other hand there can be proposed

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

as much measures s as different weights α, β, γ exist, and no obligatory idea for such weights exists. The choice depends on the context of a comparison, which is not treatable systematically. Here it is omitted to discuss combinations of weights.

But a few statements about the weights are possible. A symmetric measure requires $\beta = \gamma$. The special case of a *cost model* is included, by setting $\alpha = 0$, and also a *benefit model* can be represented by $\beta = 0$ and $\gamma = 0$.

One could criticize that our concept of location, based on sets of points ($\mathbb{R}^2$) or atoms ($\mathbb{Z}^2$), is too specific in parametrization. Indeed, other frames of (locational) reference are possible (Bittner and Stell, in press). Moreover, with the Hausdorff distance a distance measure exists which is more general in parametrization of space (Edgar, 1990). The Hausdorff distance is symmetric and one-dimensional (the set sizes above are two-dimensional in planar space). r is zero iff $A = B$. Any other value (< 0) does not allow to conclude to a topological configuration. That disadvantage cannot be adjusted because an adequate measure of common features is not known. For that reason the Hausdorff distance cannot be completed to a similarity measure.

With the binary location function (Eq. 1) only discrete regions are tested for similarity. That fits to data sets in today's spatial data bases, where a need for quality description is realized but usually not available. On the other hand, the presented model for similarity measures could be refined for uncertain or imprecise regions. The idea is to replace a binary function f by a spatial distribution function, which corresponds to a convolution of f with a distribution function, e.g. a Gaussian. The consequences have to be worked out elsewhere.

The presented similarity measures increase linear with common location. That is a consequence of setting elementary set sizes into ratios. Such a model is purely mathematical, and there is no reason to assume that cognitive concepts of humans are comparable, with the exception of simplicity.

Similarity is a general concept applied in many spatial decision problems (as well as in other disciplines). The systematic investigation succeeded by limiting to a strict frame of reference. Concentrating on *location* of two spatial objects (regions), an elementary set of similarity measures can be presented. To what extent the model can be expanded is to investigate.

Acknowledgement

The idea of this paper goes back to a discussion with Andrew Frank, and I had interesting discourses about philosophical aspects of similarity and location with Katrin Dyballa and Thomas Bittner, all Vienna.

REFERENCES

Ackermann, F., 1984. High precision digital image correlation. In: 39. Photogrammetric Week, Schriftenreihe des Instituts für Photogrammetrie, Vol. 9, Universität Stuttgart, pp. 231–244.

Bittner, T. and Stell, J., in press. A boundary-sensitive approach to qualitative location. Annals of Mathematics and Artificial Intelligence.

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.

Edgar, G. A., 1990. Measure, Topology, and Fractal Geometry. Undergraduate Texts in Mathematics, 2 edn, Springer, New York.

Egenhofer, M. J., 1997. Query processing in spatial-query-by-sketch. Journal of Visual Languages and Computing 8(4), pp. 403–424.

Egenhofer, M. J. and Al-Taha, K. K., 1992. Reasoning about gradual changes of topological relationships. In: A. U. Frank, I. Campari and U. Formentini (eds), Theories and Models of Spatio-Temporal Reasoning in Geographic Space, Springer LNCS 639, New York, pp. 196–219.

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

Egenhofer, M. J. and Herring, J. R., 1990. A mathematical framework for the definition of topological relationships. In: 4th International Symposium on Spatial Data Handling, International Geographical Union, Zürich, pp. 803–813.

Flasch, K., 1986. Das philosophische Denken im Mittelalter. Philipp Reclam jun., Stuttgart.

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.

Jähne, B., 1995. Digital Image Processing. 3 edn, Springer, Berlin.

Lakoff, G., 1987. Women, Fire, and Dangerous Things - What Categories Reveal about the Mind. The University of Chicago Press, Chicago.

Metzger, W., 1936. Gesetze des Sehens. Senckenberg-Buch, Vol. VI, W. Kramer & Co., Frankfurt am Main.

Ragia, L. and Winter, S., 1998. Contributions to a quality description of areal objects in spatial data sets. In: D. Fritsch and M. Sester (eds), ISPRS Commission IV Symposium, Stuttgart, p. submitted.

Snodgrass, R. T., 1992. Temporal databases. In: A. U. Frank, I. Campari and U. Formentini (eds), Theories and Methods of Spatio-Temporal Reasoning in Geographic Space, Lecture Notes in Computer Science, Vol. 639, Springer, Berlin, pp. 22–64.

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

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

Vosselman, G., 1992. Relational Matching. Lecture Notes in Computer Science, Vol. 628, Springer, Berlin.

Walter, V., 1997. Zuordnung von raumbezogenen Daten am Beispiel der Datenmodelle ATKIS und GDF. Phd thesis, Fakultät für Bauingenieur- und Vermessungswesen der Universität Stuttgart.

Wazinski, P., 1993. Graduated Topological Relations. Universität des Saarlandes.

Weidner, U. and Förstner, W., 1995. Towards automatic building extraction from high-resolution digital elevation models. ISPRS Journal of Photogrammetry and Remote Sensing 50(4), pp. 38–49.

Winter, S., 1996. Unsichere topologische Beziehungen zwischen ungenauen Flächen. Phd thesis, Landwirtschaftliche Fakultät der Rheinischen Friedrich-Wilhelms-Universität Bonn.

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