Cognitive Adequacy of Topological Consistency Measures*

Nieves Brisaboa[1], Miguel R. Luaces[1], and M. Andrea Rodríguez[2]

[1] Database Laboratory, University of A Coruña
Campus de Elviña, 15071 A Coruña, Spain
{brisaboa, luaces}@udc.es
[2] Universidad de Concepción, Chile
Edmundo Larenas 215, 4070409 Concepción, Chile
andrea@udec.cl

Abstract. Consistency measures provide an indication on how much a dataset satisfies a set of integrity constraints, which is useful for comparing, integrating and cleaning datasets. This work presents the notion of consistency measures and provides an evaluation of the cognitive adequacy of these measures. It evaluates the impact on the consistency measures of different parameters (overlapping size, external distance, internal distance, crossing length, and touching length) and the relative size of geometries involved in a conflict. While a human-subject testing supports our hypotheses with respect to the parameters, it rejects the significance of the relative size of geometries as a component of the consistency measures.
Keywords: topological similarity measure, inconsistency measures, spatial inconsistency

1 Introduction

A dataset is consistent if it satisfies a set of integrity constraints. These integrity constraints define valid states of the data and are usually expressed in a language that also defines the data schema (logical representation). Consistency measures provide an indication on how much a dataset satisfies a set of integrity constraints. They are useful to compare datasets and to define strategies for data cleaning and integration. Traditionally, consistency in datasets has been a binary property, the dataset is either consistent or not. At most, consistency measures count the number of elements in a dataset that violate integrity constraints, but the concept of *being partially consistent* does not exist. Spatial information rises new issues regarding the degree of consistency because the comparison of spatial data requires additional operators beyond the classical comparison operators ($=, >, <, \leq, \geq, \neq$). Geometries are typically related by topological or other spatial relations, upon which different semantic constraints may be defined.

* This work was partially funded by Fondecyt 1080138, Conicyt-Chile and by "Ministerio de Ciencia e Innovación" (PGE and FEDER) refs. TIN2009-14560-C03-02, TIN2010-21246-C02-01 and by "Xunta de Galicia (Fondos FEDER)", ref. 2010/17.

In a previous work [9], we defined a set of measures to evaluate the violation degree of spatial datasets with respect to integrity constraints that impose topological relations on the semantics of spatial objects. These measures contextualize the relative importance of the difference of the topological relation between two geometries with respect to an expected topological relation by considering the size of geometries within the whole dataset. In this paper we carry out a human-subject testing to evaluate all measures where we analyze not only the degree of violation in itself, but also the impact of the relative size of objects in the dataset as a component of the degree of violation. Three hypotheses were analyzed: (1) The four parameters used by the measures (i.e., *external distance, internal distance, crossing segment, and overlapping size*) are perceived by subjects as factors of the degree violation. (2) The *touching length* of geometries in touch, which is also not considered by the proposed measures, is not considered by subjects as a factor of the degree of violation. (3) The size of geometries involved in a conflict, with respect to other objects in the dataset, is perceived by subjects as a factor of the degree of violation.

The organization of the paper is as follows. Section 2 makes a revision of related work. In particular it analyzes different approaches to comparing topological relations. Section 3 presents preliminary concepts and consistency measures first defined in [9], while Section 4 describes the human-subject testing and its main results. Final conclusions and future research directions are given in Section 5.

2 Related Work

Related work addresses similarity measures of topological relations. Similarity measures are useful to compare the topological relation between geometries stored in a dataset with respect to an expected topological relation as expressed by an integrity constraint. We distinguish qualitative from quantitative approaches to comparing topological relations. A qualitative representation of topological relations uses a symbolic representation of spatial relations, such as the topological relations defined by Egenhofer and Franzosa [3] or by Randell *et al.* [8]. Under this representation, a similarity measure compares topological relations by the semantic distance between relations defined in a conceptual neighborhood graph [7]. The disadvantage of comparing topological relations from a qualitative perspective is that it does not make distinction between particular geometries. For example, it does not distinguish between two pairs of geometries, both disjoint, but where in one case the geometries are very close and in the other case the geometries are far apart. Even more, in most cases when semantic distance is used, all edges in the conceptual graph will usually have the same weight in the determination of the semantic distance.

A quantitative representation of topological relations is given in [1] by the distance and angle between the centroid of the objects. Using this representation, similarity between topological relations is defined as the inverse of the difference between representations. Another study [4] defines ten quantitative

measures that characterize topological relations based on metric properties, such as *length*, *area*, and *distance*. The combination of these measures gives an indication of the topological relations and their associated terms in natural language (such as *going through* and *goes up to*). The problem of using the previous measures for evaluating the degree of inconsistency is that although datasets handle geometries of objects, constraints are expressed by qualitative topological relations, and therefore, only a symbolic representation of the expected topological relations exists.

In the spatial context, only the work in [9] introduces some measures to compare the consistency of different datasets. In this previous work, given an expected topological relation between any two objects with particular semantics, a violation degree measure quantifies how different is the topological relation between the objects from the expected relation expressed by a topological constraint. While this previous work provides an evaluation with respect to semantic distance, it does not evaluate the cognitive adequacy of the measures neither the impact of the relative size of objects in the quantification of inconsistency.

3 Definition of Consistency Measures

In this work we concentrate on integrity constraints that impose topological relations depending on the semantics of objects. Particularly, we extend the *Topological Dependency* (TD) constraints defined in [2] or the semantic constraints in [5] to consider a wider range of constraints found in practise. The definitions of topological relations are those in the Open Geospatial Consortium Simple Feature Specification [6] and used in the subsequent specification of topological dependency constraints.

Let T be a topological relation, and $P(\bar{x}_1, g_1)$ and $R(\bar{x}_2, g_2)$ be predicates representing spatial entities with non-empty sequences of thematic attributes $\bar{x}_1$ and $\bar{x}_2$, and geometric attributes g_1 and g_2, respectively. A conditional topological dependency constraint is of the form:

$$\forall \bar{x}_1 \bar{x}_2 \bar{g}_1 \bar{g}_2 (P(\bar{x}_1, g_1) \wedge R(\bar{x}_2, g_2) \wedge \psi \rightarrow T(g_1', g_2'))$$

where g_1' is either g_1, $\Theta_1[g_1]$ or $\Theta_2[g_1, d]$, with d a constant and Θ_1 and Θ_2 geometric operators that return a geometry. Geometry g_2' is defined in the same way than g_1' where g_1 is replaced by g_2. Also ψ is an optional formula in conjunctive normal form (CNF) defined recursively by:

(a) $y \Delta z$ is an atomic formula, with $y \in \bar{x}_1$, $z \in \bar{y}_2$, and Δ a comparison operator $(=, \neq, >, <, \leq, \geq)$.
(b) $T'(g_1, g_2)$ is an atomic formula, with T' a topological relation.
(c) $\theta_1[g_1] \Delta c$ or $\theta_1[g_1] \Delta \theta_2[g_2]$ are atomic formula, with c a constant, Δ a comparison operator $(=, \neq, >, <, \leq, \geq)$, and θ_1, θ_2 geometric operators that return real numbers (e.g., area, length, perimeter, and so on).
(d) An atomic formula is a CNF.
(e) $(t_1 \vee t_2)$ is a clause with t_1 and t_2 atomic formulas.

(f) A clause is a CNF.
(g) $c_1 \wedge c_2$ is a CNF formula with c_1 and c_2 clauses or CNF formulas.

Using the previous definitions and considering predicate $county(idc, ids, g)$ and $state(ids, g)$, a CTD could be "a county must be within the state to which it belongs":

$$\forall idc, ids, g_1, g_2(county(idc, ids, g_1) \wedge state(ids, g_1) \rightarrow \mathsf{Within}(g_1, g_2))$$

Let ψ be an integrity constraint of the form 3 with topological relation T. A pair of tuples $P(\bar{u}_1, s_1)$ and $P(\bar{u}_2, s_2)$ is *i*nconsistent if the antecedent in ψ instantiated by $P(\bar{u}_1, s_1)$ and $P(\bar{u}_2, s_2)$ is satisfied but the consequent not. We defined in [9] a collection of measures to compute the violation degree for all topological relations between surfaces. Two main components define the degree of violation: (1) the magnitude of the conflict and (2) the relevance of the conflict. The magnitude of the conflict measures the difference between the relation held by the two geometries and the expected relation between them. For example, if two geometries must touch but they are disjoint, the magnitude of the conflict is proportional to the separation between the geometries. In the case that two geometries must touch but they overlap, the magnitude of the conflict is proportional to the overlapping area between the geometries. On the other hand, the relevance of the conflict is determined using the relative size of the objects.

We have considered five different parameters to compute the magnitude of the conflict with respect to different topological relations: (1) the *external distance* between disjoint geometries, which has an impact on conflicts risen by the separation of geometries when they must intersect (i.e., equal, touch, overlap, or within). (2) The *internal distance* between geometries when one is within the other geometry, which has an impact on conflicts when geometries must be externally connected (i.e, they must touch or they must be disjoint). (3) The *overlapping size* of geometries that are internally connected, which has an impact on conflicts when geometries must be externally connected (4) The *crossing length* that represents the length of the minimum segment of a curve that crosses another curve or a surface, which has an impact on conflicts when geometries must be externally connected. (5) The *touching length* between geometries represents the length of the common boundary between geometries. In the definition of the violation degree measures, we have used the first four parameters and we have not used the *touching length* in any of the measures.

4 Empirical Validation of Consistency Measures

To validate the cognitive adequacy of our measures, we designed human-subject tests oriented to evaluate both components of our measures: (i) the parameters that measure the magnitude of conflicts (*external distance*, *internal distance*, *overlapping size* and *crossing length*) and (ii) the computation of the conflict relevance using the relative size of the objects. We wanted to evaluate whether

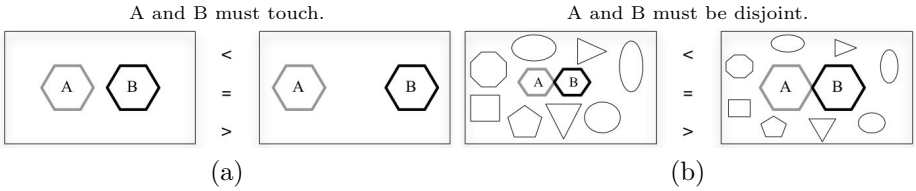

Fig. 1. Comparison of the violation degree: (a) example of section I and (b) example of section II

these parameters have an impact on the violation degree perceived by subjects and, therefore, if they are cognitively adequate to define the violation degree of CTDs. We also wanted to evaluate whether our decision to ignore the parameter *touching length* was correct, that is wheter the *touching length* has an impact on the magnitude of conflict (e.g., geometries that should touch must but are disjoint).

The following three hypotheses were studied:

H_1: *External distance*, *internal distance*, *crossing length*, and *overlapping size* are perceived and used by subjects to evaluate the degree of violation of CTDs.

H_2: *Touching length* is not considered by subjects to evaluate the degree of violation of CTDs.

H_3: The relative *size* of the geometries that participate in the violation of CTDs with respect to other objects in the dataset affects the perceived violation degree. More precisely, the larger the geometries the larger the violation degree.

The subjects of the test were 69 second year computer science students. The test was performed at the beginning of a normal class. They were not given any instructions in addition to those written in the test. Nine of the subjects were eliminated because they did not complete the test or their answers showed a clear misunderstanding of the questions. Students did not receive any compensation for answering the test, which explains why some students did not answer all questions.

The test begins with a set of figures describing the topological relations that combine geometries with different dimensions (*surfaces*, *curves* and *points*). These figures are the only reference to understand the meaning of each topological relation. Then, a short paragraph explains that the objective of the test is to evaluate the violation degree of the topological relations in the figures that follow in the next pages. The instructions emphasize that the violation of the expected topological relation was referred exclusively to the geometries clearly identified in each figure with colors blue and yellow, explaining that any other geometry coloured in black is correct.

The test consists of 3 sections. Section I includes 24 questions to check whether the parameters used by our measures are those perceived by the subjects

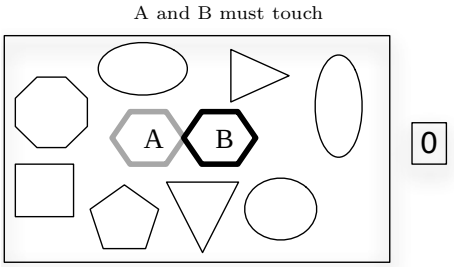

Fig. 2. Example of section III

as a factor of the violation degree (H_1 and H_2). Figure 1(a) shows a question of this section. The task is to choose whether or not one of them shows a larger ($>$), equal ($=$) or smaller ($<$) violation degree of an expected topological relation among geometries A and B. Specifically, we check that the larger the value of the parameter that defines the magnitude of conflicts in our measures, the larger the perceived violation.

The questions in this section check the four parameters considered in our measures (i.e., *external distance*, *internal distance*, *crossing length*, and *overlapping size*) as well as the influence of the *touching length* between geometries. The questions also represent a balanced selection of different topological relations and types of geometries (mainly surfaces and curves, but also points).

Section II includes 14 questions similar to those in Section I. The difference is that the figures now include black geometries that represent the context where the violation occurs. Figure 1(b) shows a question of section II. This section is designed to prove the influence of the context, that is, the influence in the perceived violation of the size of the two geometries in conflict with respect to the other geometries in the dataset (H_3).

Section III shows each single figure in the questions in section III. Fourteen of them represent small blue and yellow geometries and large context geometries and fourteen of them represent large blue and yellow geometries and small context geometries. For each figure, the subjects were asked to provide a value of the degree of violation between 0 and 100. The example given to the subjects (see Figure 2) does not violate the expected topological relation and, consequently, the assigned violation degree value is 0. We decided to use this example with value 0 to avoid any influence on the value given by subjects in case of a violation. This section is designed to validate our measures by evaluating whether or not the violation degrees computed by our measures is in concordance with the violation degrees given by the subjects. If there is a high correlation between the scores provided by the subjects and the values obtained with our measures, we can conclude that our measures are valid and that they reflect the opinions of the subjects.

We decided not to check all combinations of topological relations between geometries versus an expected topological relation because this would require 64 questions in the test. We assume that if a parameter (e.g. the *external distance* between geometries) was perceived and used to decide a degree of violation between geometries that must *touch* but are *disjoint*, then it will be also perceived and used to decide the degree of violation between two geometries that must *overlap* but are *disjoint*. Similarly, we assume that if a parameter is perceived and used for geometries of a particular dimension (e.g., surfaces), then it will be also for geometries of other dimension (e.g., curves or points). Furthermore, we did not include figures in the test where the expected relation is Equal.

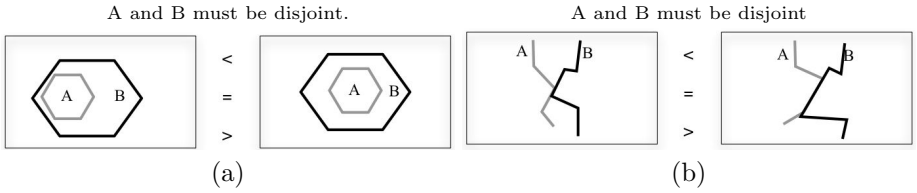

Fig. 3. Comparison of the violation degree: (a) a surface within another surface (b) a point within another point

Table 1 shows the raw data obtained in the different questions of section I. In this table, *Expected* refers to the expected topological relation as expressed by a CTD, *Actual* refers to the real relation between the geometries in the question, *Geometries* refers to the type of geometries involved in the relation, *Parameter* refers to the parameter evaluated by the test, '+' refers to the percentage of answers that perceive a positive impact of the parameter on the violation degree, '=' refers to the percentage of answers that do not perceive any effect of the parameters on the violation degree, and '−' refers to the percentage of answers that give an inverse influence of the parameter on the violation degree.

One can see that there is always around a 10% of subjects that answered a different option than the one we expected. When some of the subjects were asked about the reasons of their answers, many of them said it was a mistake. We realize now that it was difficult for the subjects to keep the level of concentration to mark correctly $<$ or $>$ on all the questions. The results prove beyond any doubt that the parameters *external distance*, *overlapped size* and *crossing length* are consistently used by the subjects to evaluate the violation degree (hypothesis H_1). We performed the Student's t_test to evaluate the significance of our results and we found that the average percentage of correct answers for the parameter *external distance* was significant at the 99% level. In the same way the results showed that for the parameters *crossing length* and *overlapping size* the average percentage we obtained is significant at the 95% level.

Results regarding the parameter *internal distance* are more difficult to analyze. For questions 2 and 23 there was a large percentage of subjects (around

Table 1. Results section I

#	Expected	Actual	Geometries	Parameter	+ %	= %	- %
1	Disjoint	Overlaps	surface × surface	Overlapping size	83	8	8
2	Disjoint	Within	surface × surface	Internal distance	48	28	23
3	Touches	Overlaps	surface × surface	Overlapping size	68	20	12
4	Touches	Within	surface × surface	Internal distance	62	22	17
5	Overlaps	Disjoint	surface × surface	External distance	87	3	10
6	Overlaps	Touches	surface × surface	Touching length	32	48	20
7	Overlaps	Within	surface × surface	Internal distance	53	40	7
8	Within	Overlaps	surface × surface	Overlapping size	68	17	15
9	Within	Touches	surface × surface	Touching length	20	65	15
10	Disjoint	Overlaps	curve × curve	Crossing length	68	23	8
11	Disjoint	Overlaps	curve × curve	Touching length	58	35	7
12	Disjoint	Overlaps	curve × curve	External distance	83	5	12
13	Disjoint	Overlaps	curve × curve	External distance	82	10	8
14	Disjoint	Overlaps	curve × curve	External distance	80	12	8
15	Disjoint	Overlaps	surface × curve	Touching length	40	52	8
16	Disjoint	Overlaps	surface × curve	Crossing length	83	8	8
17	Disjoint	Overlaps	surface × curve	Internal distance	45	45	10
18	Disjoint	Overlaps	surface × curve	Internal distance	50	28	20
19	Disjoint	Overlaps	surface × curve	External distance	87	12	7
20	Disjoint	Overlaps	surface × curve	Crossing length	68	23	8
21	Disjoint	Overlaps	surface × curve	External distance	72	18	10
22	Disjoint	Overlaps	surface × curve	Internal distance	47	43	10
23	Disjoint	Overlaps	surface × point	Internal distance	52	28	20
24	Disjoint	Overlaps	curve × point	External distance	67	27	7

40%) that did not answer as we expected. Figure 3(a) shows question number 2. When asked why they answered this way, the subjects said that geometries were more disjoint when the internal distance between them was larger. We believe that this was due to a misunderstanding of the topological relation Disjoint. Question 7 was another case where many subjects gave unexpected answers due to the misunderstanding of the topological relation Overlaps. When asked why they considered that both figures have the same degree of violation, many subjects answered that in both cases there was no violation because when a geometry is within another geometry they also overlap each other.

After eliminating questions 2, 7, and 23 where there was some misinterpretation, the Student's t-test shows that the average percentage is significant at the 85% level. This means that the *internal distance* parameter affects the perception of consistency by subjects. However, further analysis must be performed in the future to better understand why in some cases the internal distance is considered important and in some cases not.

Finally, as H_2 states, the *touching length* is not a useful parameter to evaluate the degree of violation of a topological constraint. Only question 11 shows a

higher percentage of subjects that considered the impact of *touching length* important on the violation degree. However, as it can be seen in Figure 3(b), this question was the only one where the geometries in each figure were not exactly the same. Thus, there may be factors other than *touching length* involved in the subjects' answers.

The results for Section II indicate that 35%, 35% and 30% of the subjects considered that the size of geometries in conflict had a positive, equal or negative impact on the violation degree, respectively. These results do not support our hypothesis H_3, but they also do not support the alternative hypothesis that states that the relative size has no impact or a negative impact on the violation degree. Therefore, we cannot extract any conclusion over the influence of the context in the evaluation of the violation degree. These results are in concordance with the results obtained in section III.

Finally, for each question in Section III, we computed the average score given by the 60 subjects and the value of our measure. Then, we computed the Pearson correlation between both series of values. The correlation coefficient equals to 0.54. Given that this is a very small value, we excluded the relative weight from the computation of our measures and we obtained a correlation coefficient of 0.84. This result supports the conclusion that we extracted from section II. We can conclude that the relative size of the geometries is not considered to be important by the subjects. Or at least, that the subjects consider more important the magnitude of the conflicts than the relative size of the geometries with respect to other objects in the dataset.

5 Conclusions

We obtained two types of conclusions from this work: some related to the definition of the measures and other related to the methodology to evaluate these measures. Overall, it is clear that the use of parameters such as *external distance* and *overlapping size* allows us to discriminate situations that a semantic distance approach to comparing topological relations would otherwise overlook. Unless we consider the particularity of the geometries, a pair of geometries holding the same topological relation will always have the same degree of violation with respect to a different expected topological relation.

The results of the empirical evaluation indicate that the parameters that define our measures agree with the human perception of the violation degree. The only one that was not fully confirmed was the *internal distance*, which requires further evaluation. Contrary to our expectation, the results also indicate that the relative size of geometries in conflict with respect to other geometries in the dataset has less impact on the evaluation of the violation degree than what we expected. This is confirmed by the increase in the correlation of the scores given by the users in Section III of the test when we eliminated the effect of the relative size of geometries from the measures.

We confirmed that the design of the test is critical. There are two basic problems that need to be solved for future empirical evaluations: the difficulty of the task and the knowledge the subjects need about topological relations.

Regarding the difficulty of the task, the questions in the test require a high level of concentration. This explains the high number of mistakes we found in the questions of section I. On the other hand, in section III the task was easier because only a score was requested. However, the subjects complained about the difficulty and many of them moved back and forward changing the scores while answering the questions.

The problem of the knowledge about the topological relations is harder to solve. Explaining the meaning of the topological relations before the test does not guarantee that they use these definitions instead of their own interpretations. For instance, some of the subjects considered that two surfaces that *overlap*, also *touch*, or that two surfaces that are one *within* the other also *overlap*. The only way to avoid this problem is to train the subjects in the meaning of the topological relations. However, it may be difficult to do this without instructing them in our view of the parameters that define the measure of the violation degree. Probably, the safest way to tackle this problem is to select the figures and their relations very carefully to avoid that subjects misunderstand topological relations.

References

1. Berreti, S., Bimbo, A.D., Vicario, E.: The computational aspect of retrieval by spatial arrangement. In: Intl. Conference on Pattern Recognition (2000)
2. Bravo, L., Rodríguez, M.A.: Semantic integrity constraints for spatial databases. In: Proc. of the 3rd Alberto Mendelzon Intl. Workshop on Foundations of Data Management, Arequipa, Peru. vol. 450 (2009)
3. Egenhofer, M., Franzosa, R.: Point Set Topological Relations. IJGIS 5, 161–174 (1991)
4. Egenhofer, M., Shariff, A.: Metric details for natural-language spatial relations. ACM Transactions on Information Systems 16(4), 295–321 (1998)
5. Hadzilacos, T., Tryfona, N.: A Model for Expressing Topological Integrity Constraints in Geographic Databases. In: Spatio-Temporal Reasoning. pp. 252–268. Springer LNCS 639 (1992)
6. OpenGis: Opengis Simple Features Specification for SQL. Tech. rep., Open GIS Consortium (1999)
7. Papadias, D., Mamoulis, N., Delis, V.: Algorithms for querying spatial structure. In: VLDB Conference. pp. 546–557 (1998)
8. Randell, D., Cui, Z., Cohn, A.: A spatial logic based on regions and connection. In: Nebel, B., Rich, C., Swarthout, W. (eds.) Principles of Knowledge Representation and Reasoning. pp. 165–176. Morgan Kaufmann (1992)
9. Rodríguez, M.A., Brisaboa, N.R., Meza, J., Luaces, M.R.: Measuring consistency with respect to topological dependency constraints. In: 18th ACM SIGSPATIAL Intl. Symposium on Advances in Geographic Information Systems, ACM-GIS 2010, San Jose, CA, USA. pp. 182–191 (2010)