Exploiting Spatial Vagueness in Spatial OLAP: Towards a New Hybrid Risk-Aware Design Approach

Elodie Edoh-Alove
Sandro Bimonte
François Pinet
TSCF, Irstea
24 Av. des Landais, 63172, Aubière, France
{name.surname}@irstea.fr

Yvan Bédard
Department of Geomatics Sciences and
Centre for Research in Geomatics,
Laval University
Quebec City, Quebec, Canada
yvan.bedard@scg.ulaval.ca

Abstract

Spatial OLAP systems allow multidimensional analysis of huge volume of spatial data. Spatial vagueness is one of the most transparent and current imperfections of spatial data. Although several works propose new ad-hoc models for handling spatial vagueness in Spatial DBMS, the implementation of those models in conventional systems and Spatial Data Warehouses (SDW) is still in an embryonic state. Thus, to help reduce that uncertainty and its consequences in SOLAP analysis, we present a new approach for designing spatial data cubes based on users tolerance to the risk of misusing datacubes.
Keywords: Spatial vagueness, Risk of use, Spatial OLAP.

1 Introduction

Spatial OLAP (SOLAP) can be defined as a visual platform built especially to support rapid and easy spatiotemporal analysis and exploration of data following a multidimensional approach comprised of aggregation levels available in cartographic displays as well as in tabular and diagram displays. It generally allows the mapping, comprehension and comparison of the geographic distribution of studied phenomena. The explored data are stored in a Spatial Data Warehouse (SDW) as datacube, that is the multidimensional model implementation[16]. Usual clients of those technologies are first and foremost decision-makers who are rarely fully aware of the problems related to spatial data uncertainty. And yet, spatial data are primarily "false" but useful models of the reality [6] since they define geographic objects by means of crisp boundaries even if it is not always possible to define exactly when objects begin and end (e.g. Ocean) (spatial vagueness). With such a choice of representation, a clear gap is created between majority of real world phenomena and their formal representation in spatial databases [5].

On the other hand, SDW design is a very crucial step since the spatio-multidimensional model defines also SOLAP operations allowed for decision-makers. In that line, only [13] advocate for a methodology for designing SDW based on the data driven approach. Indeed, to present, different approaches can be used to design (spatial) DW: user-driven approach, data-driven approach and hybrid approach [13]. The classic result of design methods implementing one of these approaches is the multidimensional schema of a SDW which is fed with facts computed using the available data sources. However, these approaches do not take into account quality of spatial data (and in particular spatial vagueness), which can lead to incorrect analysis. As an illustration, let us consider an agricultural SOLAP application to analyze the number of plots in France that are within the regulated distance from a watercourse, to help make the decision to conduct regulation actions on some plots in order to prevent potential water pollution (e.g. contamination by pesticides). In reality, those watercourses have a minimum extent (extent during dry periods) and a maximum extent (extent during periods of rise in the water level). Thus, a crisp representation of these spatial objects conducts to erroneous analysis (facts).

To best of our knowledge only some recent works introduce spatial vagueness in the spatio-multidimensional model [18] extending it, without proposing any efficient implementation.

Thus in this position paper, motivated by the relevance of exploiting spatial data marked with spatial vagueness in existing SOLAP systems, we propose the main outlines of a risk-aware SDW design approach (section 3). But first, we also present a state of art of the existing work (section 2).

2 Related work

2.1 Spatial vagueness and data warehouse design

Geomatics community uses the term "vagueness" to characterize a geographic object for which it is not possible to define the spatial extent. For those objects, that are said to be ill-defined, no properties combination allows unequivocal identification and allocation in a discrete class or even the definition of their precise spatial extent. Vagueness here is not only a matter of shape and position, but also a matter of vague conception (e.g. what do we call ocean?) of the entity, fuzziness in its identification (e.g. when a tree is a tree?) or descriptive qualitative attributes (poor or rich soil) [1].

In computer science community, the term "spatial vagueness" is used to describe either (1) "fuzziness" (fuzzy boundaries) or (2) "uncertainty" (position or measurement uncertainty) [17]. Fuzziness here is an inherent property of a spatial phenomenon that certainly has an extent but does not have well-defined boundaries (Ex: flood zones). That notion will be generalized recently by [2] to the notion of "shape vagueness" which is then defined as the imperfection on the

shape of some geographic objects. Shape vagueness is viewed as a matter of fuzzy boundaries or in a more general way, broad shape. Another term used in the community for uncertainty (2) is "location vagueness" which is either a matter of lack of knowledge about the position and shape of an object with an existing real boundary (position uncertainty) or the inability of measuring such an object precisely (measure uncertainty)[9, 17].

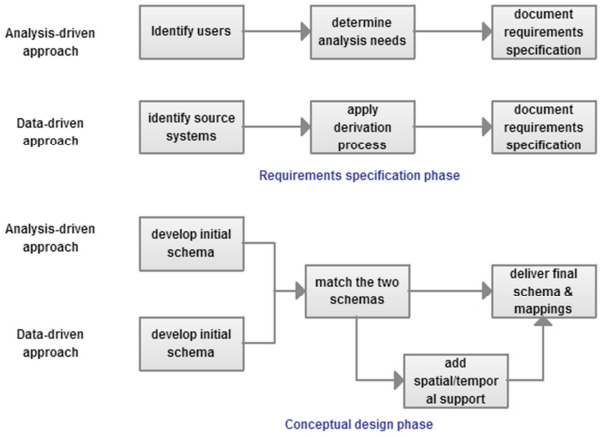
In our work we want to focus on shape and location vagueness and their influences on other derived spatial data that could be stored in the datacube (e.g. surface of flood zones, distance from a watercourse etc.). The spatial data that have broad shape or vague location plus their derived data will be designated in this paper as spatial data marked with spatial vagueness. Note that we view spatial data, according to the international standards in Geomatics, as any data (e.g. address, shape etc.) that is used to localize spatial features in the geographic space.

In order to reduce uncertainty related to spatial vagueness, in conventional systems (spatial databases, GIS etc.), researchers, throughout the years, have focused on the use of vague objects (as opposed to crisp objects) to represent some spatial phenomena in a more accurate way. Thus, four alternatives to the crisp object type (point, line, polyline) have been found: exacts models [2] where the geographic information is represented by a complex geometry consisting of at least two crisp geometries (minimum extent areas where the phenomenon is surely present and maximum extent areas where the phenomenon is probably present), fuzzy models (based on the Fuzzy Set theory) that describes the possibility that an individual is a member of a set or that a statement is true, probabilistic models [4] and rough models [20].

Note that despite the effort, to present, the majority of GIS and spatial databases still manage only crisp geometries[14].

Regarding data warehouses projects development, its phases are generally: the requirements specification (the needs are analysed), the conceptual-design (conceptual and logical schemas are produced), the implementation (a physical schema is produced), and the feeding (Extract Transform and Load-ETL operations are done). There are numerous approaches proposed in both computer science and Geomatics communities to develop a data warehouse. They differ essentially on the approaches used for specifying requirements, which allows us to classified them in [13]: (1)the user-driven approach where the user has an important role to play in defining the specifications of the data warehouse project, (2) the business-driven approach where business requirements or processes are analysed in order to deduce facts and dimensions (e.g. considering the metrics used by decision makers, to evaluate business activities, as measures), (3) the source-driven approach where the data sources are exploited in a semi-automatic/automatic way to extract facts and dimensions and finally (4) the hybrid approach where user/business driven requirements specification activities are usually conducted in parallel with the source-driven ones, leading to the creation of two data warehouse schemas that will be matched afterwards to obtain the final schema . User-driven and business-driven are often classified under the same category named goal-driven or analysis-driven [13]; also, for spatial data warehouses in specific, only [13] introduced data-driven and hybrid (Figure 1) approaches. Their main contribution is specifying that the spatial/temporal support must be added (if non-existent in the sources) before the final schema delivery. Among the existing design methods implementing those approaches [15], some address the problem of quality from the logical inconsistency aspect by proposing definition of integrity constraints to ensure the consistency of multidimensional models or data[3], others from the quality of aggregation operations aspect [8]. None has specifically addressed spatial vagueness.

Figure 1: Steps of the hybrid (analysis/data-driven) approach for spatial and temporal data warehouses [13]



2.2 Spatial vagueness management in data warehouses

The practical integration of vague objects in the data warehouse is in an embryonic state and it is only very recently that [11] have proposed an algorithmic approach based on Fuzzy Set theory to deal with the problem of fuzzy boundaries of erosion risk areas among others. On their side [18] have redefined the multidimensional model in order to take into account exact models, especially those [2, 14] have worked out. They introduced the term "vague" in multidimensional concepts, redefining spatial attributes, measures, dimensions and hierarchies. For example, "vague spatial attribute" is a spatial attribute that holds vague spatial objects (a vector composed of a pair of crisp spatial objects, namely the core and the dubiety) in its domain. There are no implementation tools proposed with their new definitions.

Thus, even though integrating vague objects in data warehouses is a good approach to reduce the uncertainty related to spatial vagueness, there is still much to do in order to design, implement and exploit spatial data warehouses with vague objects. As a matter of fact, existing tools (ETL, SOLAP server and client) and databases [14] are not designed to manage vague objects. In addition, this type of solution does not address where spatial members or measures stored in the data warehouse do not have geometries though non-geometric spatial data are common in the majority of spatial data cubes[16]. Yet it might be the best solution for SOLAP applications where the risk of use when exploiting data marked with spatial vagueness in the analyses is absolutely unacceptable no matter what.

For other application cases where different degree of risk of use can be tolerated depending on the SDW intended use, what possibilities producers have?

3 Hybrid risk aware approach for spatial datacubes design

Our main objective is to work out an approach that would help spatial datacube producers to design datacubes while taking into account, not only users' needs and available data sources, but also the risks users are able to tolerate, relatively to spatial vagueness, for requirements specification and for the final datacube schema production. In other terms, we want to propose a new approach for designing SDW based on the concept of risk related to spatial vagueness in data cubes.

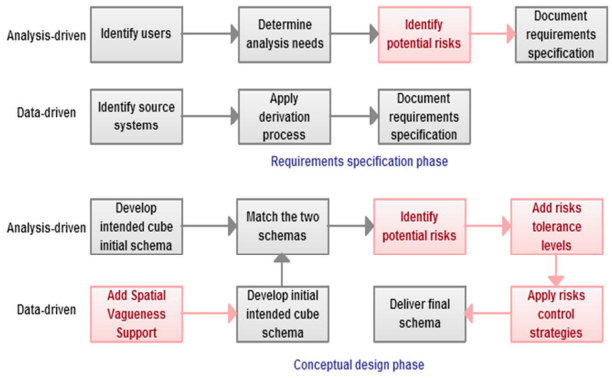
The common definition adopted in Geomatics field is the one proposed by [10]: "*A risk is a combination of the probability of occurrence of a harm and the severity of that harm*". Therefore, in a SOLAP systems context, a risk of use is the risk related to the action of exploiting or/and interpreting data in a decisional process. Note that harm, in the same context, refer to an inappropriate use of a datacube [12].

To best of our knowledge, only [12] and [7] investigate risk for SOLAP applications. They propose a risk management method whose phases are: identify possible inappropriate use of spatial datacube (identify risks), identify and establish strategies for those misuses prevention (risk analysis and evaluation plus responses toward the risks), monitor the risks, and finally document the risk-management process. Their preliminary work provide a first classification of risks related to all spatial data quality issues in data cubes, and some visual policy to prevent users. However, they do not give the tools to deduce a specific data cube (spatio-multidimensional schema and aggregations) based on acceptable risks in a systematic hybrid approach. Moreover, they do not use a formalism to define risks and quality issues, which is a mandatory issue for datacube design approaches.

Let us introduce our proposal considering the example of Section 1. The risk of misuse comes essentially from watercourses and derived computed data such as distance between plots and watercourses. We know that the decision-maker want to analyze the number of plots, plus their localization, that are within the regulated distance from watercourses. In this case, the potential risk related to the intended spatial datacube use is quite high and cannot be tolerated by the analyst. Therefore, our approach will propose, for example, to compute the distance by using the maximal extent of watercourses in order to reduce that risk (*spatial data cube regulation*). For another analyst who is only interested in analyzing budget spent on plots agricultural activities, that risk will be non-existent (*data cube budget*). Thus, our approach might propose to ignore watercourses vagueness and produce a classic spatial datacube.

The steps (for requirements specification and conceptual-design) of our proposed approach, extending the classical approach (Figure 1), are presented in Figure 2. In particular, in the requirements specification process, designer, with the help of identified users, define potential data cube elements (facts, dimensions and aggregations) and the related risks (*Identified potential risks*) (e.g. a risk on the "distance" facts). At the same time, we extend the conceptual design approach with a step (*Add Spatial Vagueness Support*) that allows to retrieve (i.e. mark) data with their spatial vagueness (e.g. for each watercourse the min and max geometric extensions are retrieved if existent). After that, the datacube schemas drawn from analysis-driven and data-driven approaches are matched (*Match the two schemas*). The identified risks list is updated at this point and then, the risks are organized (*Add risks tolerance levels*) in tolerance levels (*spatial data cube regulation* with tolerance level 1 (Low) and *data cube budget* with a tolerance level 2 (High)) with corresponding management strategy (particular distance computing (Low) and indifference (High)). The next step in the approach is the application of the strategies (*Apply risks control strategies*), depending on the tolerance level indicated by a particular user, in order to extract the final appropriate schema (e.g. in our case study, for the budget manager for example, only the *data cube budget* is provided).

Figure 2: Hybrid risk-aware Approach



In our future work, we will elaborate conceptual-design tools (formalisms etc.) and semi-automatic method by extending existing ones [3, 12, 15]. Let us note some particularities of our method:

(1) It uses traditional data sources (with crisp objects): in fact, the data sources are all the available transactional systems in the organization since the whole purpose of the data warehouse is to help extract new knowledge from the existing. This allows implementing generated spatial data cubes in any existing SOLAP tool.

(2) It provides schemas, aggregations and visualization elements: the classic hybrid methods only provide schemas (multidimensional, logical or physical) as outputs, but our method will also specify the different pertinent and authorized aggregation operations, as well as visualization elements. For example, when aggregating on coarser spatial level (e.g. department) the max aggregation function should consider only measures associated with watercourses having accurate geometry (no spatial vagueness). About visualization, as already stated in [12], it is mandatory to use ad-hoc visual policies (e.g. red pivot table cells) to inform user that some measure values are associated to watercourses with spatial vagueness issues.

(3) It is semi-automatic: the method allows automatic extractions of schemas (with mention of the aggregation

operations and visualization elements) from data sources and/or users' requirements. The designer can interact with it in order to validate the schemas. Indeed, in complex SOLAP applications, such as agro-environmental ones, users need several spatial data cubes prototypes to achieve the satisfying one. To do so, we will base our method on standard design languages, such as UML, ER, since several existing work [3] have already proven that they simplify, accelerate and allow spatial datacube conception and development process automation.

4 Conclusion

In this paper, we present a review of the existing works on spatial vagueness management in SOLAP systems. We noticed that the risk management perspective is a more interesting way to guarantee such reliability without asking too much extra efforts from the users. Thereby, we proposed steps for a risk-aware hybrid approach for designing spatial datacubes. We are currently working on the implementation of our approach extending/integrating [3, 12, 15]. We will validate our method using French spread data [19]

References

[1] Bédard Yvan. A study of Data using a Communication based Conceptual Framework of land Information Systems. *Le Géomètre canadien*, 40:12, 1986.

[2] Bejaoui Lotfi *Qualitative topological relationships for objects with possibly vague shapes: implications on the specification of topological integrity constraints in transactional spatial databases and in spatial data warehouses*. Université Blaise Pascal, 2009.

[3] Boulil Kamal, Bimonte Sandro , Pinet Francois, Chanet Jean-Pierre A UML Profile and OCL-based Constraints for Spatial Data Cubes. *Information Systems*:45, 2011.

[4] Burrough Peter A., Frank Andrew U. *Geographic Objects with Indeterminate Boundaries* London, 1996.

[5] Cheng Tao, Molenaar Martien, Lin Hui. Formalizing fuzzy objects from uncertain classification results. *International Journal of Geographical Information Science*, 15:27-42, 2001.

[6] Devillers Rodolphe, Jeansoulin Robert. *Qualité de l'information géographique*. Paris, 2005.

[7] Gervais Marc, Bédard Yvan, Levesque Marie-Andrée, Bernier Eveline, Devillers Rodolphe. Data Quality Issues and Geographic Knowledge Discovery. *Geographic Data Mining and Knowledge Discovery*:99-115, 2009.

[8] Grenier Eve, Bédard Yvan, Chrisman Nicholas. L'agrégation de données géodécisionnelles: questions pour mieux la définir. Géomatique 2011. Montréal, Canada.p., 2011.

[9] Hazarika Shyamanta, Cohn Anthony. Qualitative Spatio-Temporal Continuity Spatial Information Theory. In Montello D., editor., p. 92-107, 2001.

[10] ISO/IEC. ISO/IEC 51 Safety aspects - Guidelines for their inclusion in standards.p., 1999.

[11] Jadidi Amaneh, Mostafavi Mir Abolfazi, Bédard Yvan, Long Bernard. Towards an Integrated Spatial Decision Support System to Improve Coastal Erosion Risk Assessment: Modeling and Representation of Risk Zones. FIG Working Week 2012. Rome, Italy.p. 6-10, 2012.

[12] Lévesque Marie-Andrée. *Formal Approach for a better identification and management of risks of inappropriate use of geodecisional data*. Laval University, 2008.

[13] Malinowski Elzbieta, Zimányi Esteban. *Advanced Data Warehouse Design : From Conventional to Spatial and Temporal Applications*. Berlin, 2008.

[14] Pauly Alejandro, Schneider Markus. VASA: An algebra for vague spatial data in databases. *Information Systems*, 35:111-138, 2010.

[15] Romero Oscar, Abelló Alberto. A Survey of Multidimensional Modeling Methodologies.p. 1-23, 2009.

[16] Salehi Mehrdad, Bédard Yvan, Rivest Sonia. A Formal Conceptual Model and Definitional Framework for Spatial Datacubes. *Geomatica*, 64:313-326, 2010.

[17] Schneider Markus. Uncertainty Management for Spatial Datain Databases: Fuzzy Spatial Data Types Advances in Spatial Databases. In Güting R., Papadias D., Lochovsky F., editors., p. 330-351, 1999.

[18] Siqueira Thiago Luís Lopes, Aguiar Ciferri Cristina Dutra, Times Valéria Cesário, Ciferri Ricardo Rodrigues. Towards Vague Geographic Data Warehouses. In Xiao N., Kwan M.-P., Goodchild M., Shekhar S., editors. *Geographic Information Science*, p. 173-186, 2012.

[19] Soulignac V., Barnabé F., Rat D., David F. SIGEMO: un système d'information pour la gestion des épandages de matières organiques - Du cahier de charges à l'outil opérationnel. *Ingénieries*:37-42, 2006.

[20] Worboys Michael. Computation with imprecise geospatial data. *Computers, Environment and Urban Systems*, 22:85-106, 1998.