Mathematical support for combining geospatial data

Boris Kovalerchuk and James Schwing
Department of Computer Science
Central Washington University
Ellensburg, WA 98926-7520, USA

Abstract

This paper describes a task-driven framework for characterizing the quality of conflated data relative to a given problem by addressing mathematical issues. Issues considered include: growing geospatial data from multiple sources, a variety of techniques for data generation, a variety of requested data combinations, and emerging data types. Briefly our approach is task-driven includes the development of task-specific measures, the use of a task-driven conflation agent, and the identification of task-related default parameters. We develop a methodology for characterizing the quality of conflated data relative to a task-specific problem statement by introducing: measures of the correctness for the combination of geospatial data using the formalized concepts of context space that integrate fuzzy logic measure along with the concept of visual correlation formalism.

1. Introduction

The aim of this paper is to consider the mathematical support to geospatial data conflation through the development of a task-driven framework for characterizing the quality of conflated data relative to a given problem. Such issues include: growing geospatial data from multiple sources, a variety of techniques for data generation, a variety of requested data combinations, and emerging data types. We concentrate on improving geospatial data representations by investigating methodologies and algorithms for conflation of disparate elevations, features, and image data sets.

The paper is organized as follows: Introduction covers background, objectives, specifics and novelty of the approach, and source data; Section 2 reviews the state of the art; Section 3 presents examples that illustrate the application of our methods along with some of the details; and Section 4 provides our references.

1.1 Background

It is well known that computers and mathematical methods have had a profound impact upon cartography. There has been a massive proliferation of spatial data, and no longer is the traditional paper map the final product. In fact, the focus of cartography has shifted from map production to the management, combination, and presentation of spatial data. Maps can be (and often are) produced on-demand for any number of specialized purposes. Unfortunately, data are not always consistent. As such, data combination (or conflation) has become a significant issue in cartography.

In a computer environment, almost any two datasets can be combined – in hopes that the end result is "better" than the sum of the initial datasets. This is not always the case, as the appropriateness of conflation depends upon two primary factors: the quality of the input data and the task-specific goal of the combination process itself. Further, the resulting information is frequently used in an environment where timely response is critical.

Consider a simple example. Suppose that a task-specific goal may be to locate individual buildings at a spatial accuracy of ±20 meters. Suppose further that there are two spatial datasets available – one set with roads at ±5 meters and the other set containing both roads and buildings at ±50 meters. Obviously, neither dataset can properly answer the question. However, if the two datasets are conflated, is the spatial accuracy of the new image ±20 meters or better? If so, the process is a success. If not, the users will either have to find new data or just accept the inaccuracies in one dataset. As a side

note, it is important that the conflation process should not be used to simplify datasets (i.e., – combine two datasets into one and delete the original data), but rather to answer specific questions.

Our overall goals include: (i) combining geospatial data, measuring conflict in the combined data, (ii) deconflicting the combined data, and (iii) testing the appropriateness of the conflation relative to the stated problem definition. Further, we see this as part of a framework that integrates our proposed mathematical measures and processes with automatic rule-based parameter adjustment and visualized results. Figure 1 illustrates the framework for the overall process.

In Figure 1, we show the relationship between three systems:

1. System of task-specific measures of correctness of conflation,
2. System of task-specific conflation methods, and
3. System of visualization and visual correlation tools.

We assume that there is a database of task-specific measures of correctness of conflation and a database of task-specific conflation methods. A specific task, *Task A*, is matched to each database and a task-specific measure and method are retrieved. Then the conflation method is applied to *Task A* and the result is tested using the task-specific measure of correctness. If the result is appropriate for *Task A* then it is visualized and delivered to the end user. Otherwise, the parameters of the conflation method are modified and the procedure is repeated until an acceptable level is achieved.

The current paper addresses the challenges of designing the first system along with the mathematical aspects of conflation methods for the second system and its interaction with the first system.

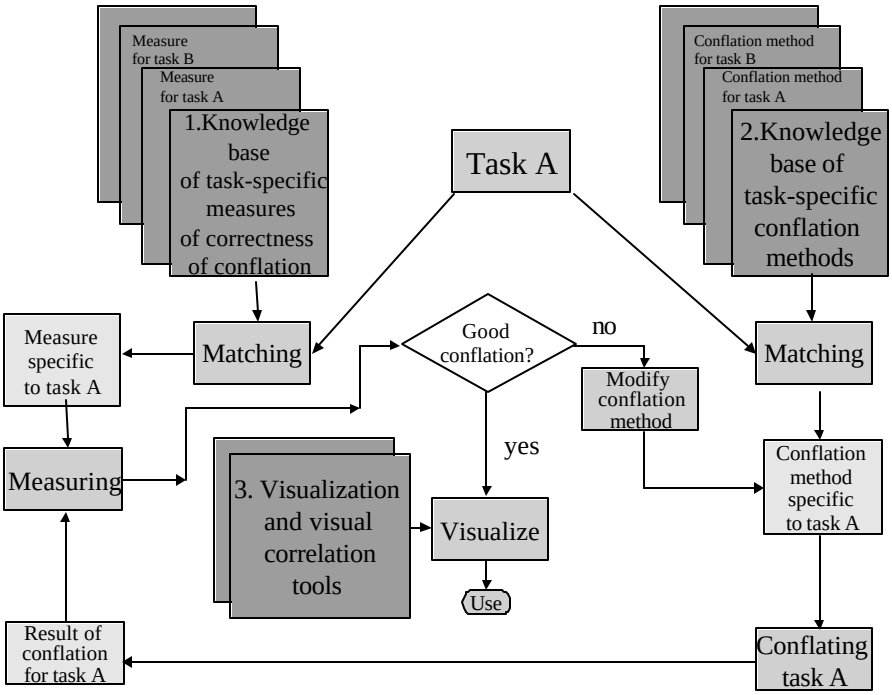


Figure 1. Framework for the overall process.

1.2 Objectives and novelty of approach

Our objective is to develop a conceptual task-driven approach mathematical framework for geospatial data conflation. This approach is intended to characterize the quality of conflated data relative to a given task.

The novelty of the approach includes:

- task-specific measures,
- use of a task-driven conflation agent, and
- identification of task-related default parameters.
- The use of visual correlation tools.
- The use of region-based complexity measures.
- The use of context space measures.

Methodology is proposed for two complimentary situations: extensive auxiliary quality information is available and little or no auxiliary quality information is available.

1.3 Source data

For enhancing data conflation process the following metadata are considered:

- Data-based statistical quality information such as random error and bias characteristics of digital terrain elevation data, and location error for a feature and
- Expert-based quality information such as fuzzy sets and linguistic terms and variables such as "topologically clean", "well matched", "highly contradictory", "bias", "consistent around polygons", "noticeable" (noticeable edge breaks of approximately 1 to 3 vertical units of resolution).

We define data quality as the accuracy, precision, completeness, and consistency of spatial data over space, time, and theme. It is assessed relative to the database specifications (i.e., – if the database specification states that objects must be located within ±100 meters, and all objects are located to that accuracy, the data is 100% accurate). As such, appropriate use (or combination) of data is always relative to both the desired output accuracy/specifications and the quality of the input data.

Three important classes of spatial objects are under consideration:

- The geometry only class, which assumes drawing, display, and geometrically defined operations.
- The geometry and topology class, which assumes vector data structures that use geometric drawing and topological operations.
- The topology only class, which assumes certain analytical operations.

2. Review

2.1 Challenges in measure design

Conflation of data includes steps such as:

- Identifying (matching) local data units to be conflated.
- Determining/selecting the "best" conflation methods of two objects

A single common flexible framework is needed that will integrate diverse types of spatial data with the following capabilities [5]:

- horizontal data integration (merging adjacent data sets),
- vertical data integration (operations involving the overlaying of maps),
- temporal data integration.
- handling differences in data content, scales, methods, standards, definitions, practices,
- managing uncertainty and representation differences,
- detecting and deal with redundancy and ambiguity of representation,
- keeping some items unmatched,
- keeping some items to be matched with limited confidence.

Currently the spatial positional accuracy (x,y,z) of a geographic observation is evaluated using root-mean-square-error (RMSE) statistics or circle of uncertainty. Current challenges include: measuring the statistical accuracy of map products derived from multiple dates of analysis, measuring the accuracy of a change

detection map, and comparing the radiometric characteristics on two anniversary dates of remotely sensed data.

In [6], the author presented several challenges, which we formulated as follows:

The **representational challenge –** to find the way of merging spatial data from variety of sources without contradiction. Often this challenge cannot be fully met.

The **uncertainty challenge** – to find a way of measuring and modeling, and summarizing inconsistencies in merged data. Often inconsistencies are inevitable in merging spatial data.

The **visualization challenge** – to find a way to visualize differences between a different digital representations and real phenomena.

2.2 Measures of correctness of conflation

Next let us review and analyze measures of correctness used in variety of fields to find a common ground for new geospatial applications. Systems can be classified in a way that varies in their level of exact definition and presentation of measures of closeness and their confidence as follows:

- *High level.* A system makes exact and clear measure of closeness on the design stage.
- *Medium level.* A system does not measure closeness of spatial objects in advance, but provides a user with interactive tools such as the curve-matching cursor.
- *Low level.* A system mostly relies on a informal human perceptual measuring mechanism, providing similar graphical presentation of entities and some pointing mechanisms between them.

A major concern with the first approach (referred to as the high level above) is in the nature of a measure of closeness as a single numeric indicator. If we try to catch the closeness of data sets with 1000 points in each of them, this measure may capture:

- *average closeness* – the averaged distance between entities using some formal definition of the measure of closeness between individual points,
- *optimistic closeness* – a measure with higher weights on smaller discrepancies in distances between entities using some formal definition of the measure of closeness between individual points,
- *pessimistic closeness* – a measure with higher weights on larger discrepancies in distances between entities using some formal definition of the measure of closeness between individual points.

Probability theory, mathematical statistics, functional analysis, fuzzy logic, machine learning and pattern recognition provide plenty of examples of measures for all three alternatives and many intermediate measures between them. The problem is that each of them hides/ignores some of the distortions, which may be critical for a specific task. A custom design of measures of closeness for a specific task and spatial entities falls into another trap -- loss of the universality of the approach and tools. In this case, there is also the need for highly task-specialized (unique) measure designs.

We suggest to extent this rule-based approach with a task-driven conflation approach:

- *Rules for matching local units and selecting a conflation method should be derived from the user's goals and tested against the user's goal.*
- *If the user has multiple goals then a separate alternative rule-sets should be generated for each goal and kept in the knowledge base.*
- *If the user is uncertain about goals of use of the conflated map some heuristic rules should be generated and tested for randomly selected potential goals.*

We suggest to implement this approach introducing the task-driven intelligent conflation agents extending the concept of a single intelligent conflation agent. These agents should detect multiple feature representations and implement conflict resolution strategies according to the goal. For instance, if the task is a global strategic overview of country's

conditions by an analyst then strategic conflation agent is activated. If the goal is to support a local reconnaissance unit then the system monitor should activate a specialized local reconnaissance conflation agent.

Thus, we argue for the dynamic selection of conflation agents vs. a static approach with a single conflation agent. The user will get a conflated map from a specialized conflation agent depending on the current task. In our approach, the customization is also a task-driven. The system monitor controls map discrepancies related to a specific user's task, such as strategic overview or local reconnaissance.

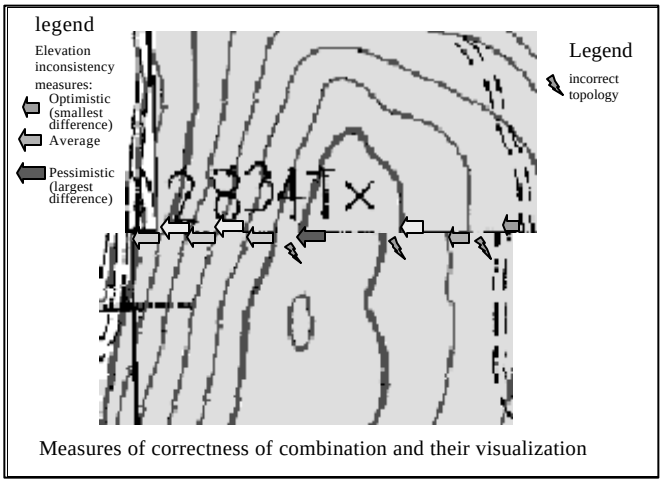
This prototype uses interactive dynamic set up of conflation characteristics by a user. The set of menus allow user: to declare the conflation tools that will be applied, to declare parameters for conflation (all tools active), to reset or recalculate the parameters for conflation, to select method for determining matched features, and to select method for evaluating links.

3. The approach

We offer multidimensional measures of correctness of conflation and their visual presentation. Figures 2 illustrates a 3-dimensional measure with three components:
(1) optimistic (short arrow), (2) average (middle arrow) and (3) pessimistic (large arrow). We also utilize features such as: location, rotation, blinking, and lighting to present contradictory characteristics of conflated maps. Figure 2 visualizes a conflation procedure, showing directions of movement and matching lines for moving the upper map to conflate two maps. This visualization also can serve as a guide in manual visual conflating, when for instance, an automatic conflation fails.

Figure 2 uses the symbol to portray contradictory breaks in elevation lines. This is an example of our matching attribute approach: if two objects a and b contain contradictory data then a is associated with the attribute "Contradicts b" and b is associated with the attribute "Contradicts a", both attributes are portrayed by the same symbol with the same color and this symbol is attached to the both objects. Similarly, if two other objects contradict each other, then the same symbol is attached to both objects.

Figure 2. Multidimensional measures of correctness



To distinguish two pairs of contradictory objects the symbol another color is used for paired objects. Thus, matching attributes support visual correlation between contradictory objects. For portraying the magnitude of contradiction between objects/information, we plan to use measures of correctness of conflation (including fuzzy logic measures) to be developed in this paper. The symbol of contradiction is rotated in accordance with this measure value. The larger rotation angle (up to 180^{o}) means the larger contradiction.

The core of our approach is to expand conceptual model of spatial data by developing a metamodel for conflating conceptual models for pairs of spatial entities. Below we briefly outline this concept. Our approach to the problem of geospatial Conflict Detection and Resolution is an extension of current rule-based approach in this area.

We enhance the rule-based approach using linguistic fuzzy logic rules, which have proved successful in many applications including fuzzy control. This allows us to accommodate both

flexibility and uncertainty in the use of linguistic terms in a quantitative manner for improving geospatial conflict resolution. These rules address conflict detection and resolution for locational attributes, nonlocational attributes.

We call a conflation procedure completely consistent if a resulting combined map: preserves relative distances between elevation lines and objects on each map preserves absolute elevations and locations on the border connecting maps, and avoids discontinuity of the objects on the border.

In the situation when completely consistent conflation is impossible, some non-linear distorting conflation methods are used as part of USGS standard. These methods differ in the number of neighboring elevation profiles involved in interpolation. The number of profiles depends on the number of vertical resolution units where the edge breaks. For such situations we design measures of correctness of conflation as compositions of: (1) measuring the distortion of relative distances on both maps,
(2) measuring the distortion of absolute elevations and locations of objects on the edge and on the interpolated profiles, (3) measuring the discontinuity on the border, and
(4) measuring the distortion of topology of objects due to composing two maps. We develop all four measures in three forms: pessimistic, optimistic and averaging (see above) relationships. A separation distance between parts of the spatial object is modeled as a fuzzy set or rough set instead of a traditional crisp set statements.

We conclude our description with a look at task-driven fuzzy logic measures. The standard vertical root-mean-square (RMSE) error statistic used in DEM is an example of an average closeness measure. The USGS provides values for standard measures such as the 7.5-minute Digital Elevation Model. Figures 3a and 3b translate some of these data into a fuzzy logic format. For example, a first membership function might represent the *desired* RMSE. Similarly, second and third membership functions data *temporarily retained* and data *rejected* which represent less desirable and unacceptable values of RSME respectively. Figure 3a may better fit practice when there is not much difference between 6.9m and 7.0m RMSE in contrast with sharp border of less than 7.0m.

Figure 3b shows a more flexible version of these functions with wider sets of uncertain values between desired, retained temporarily and rejected values of RMSE, which can be more realistic measures in some tasks. We propose to using this type of functions to match a specific task with available data accuracy. Terms "desired"(for the task), "retain temporarily" (for the task), and "reject" (for the task) are characteristics of the task and different tasks may have different "desire" and their definitions, as we illustrated in Figures 3a and 3b. The value of RMSE is a property of the data and exists independently of the task. The novelty of the approach is that we consistently distinguish properties of data and properties of the task. Hierarchical fuzzy logic membership functions form the basis of our context space concept.

4. References

[1] Chung, M., M. Cobb, K. Shaw, D. Arctur, "An Object-Oriented Approach for Handling U.S. Defense Mapping Agency Vector Product Format Databases," Proc. ISPRS WG II on New Dev. in Geophysics Information Syst., Milan, Italy, March 1996.

[2] Cobb MA, Petry FE, Shaw KB, Fuzzy spatial relationship refinements based on minimum bounding rectangle variations, Fuzzy sets and systems, 113: (1) 111-120,Jul., 2000

[3] Ehlschlaeger C. Representing uncertainty of area class maps with a correlated inter-map cell swapping heuristic, 2000, http://everest.hunter.cuny.edu/~chuck/urban/urban.html

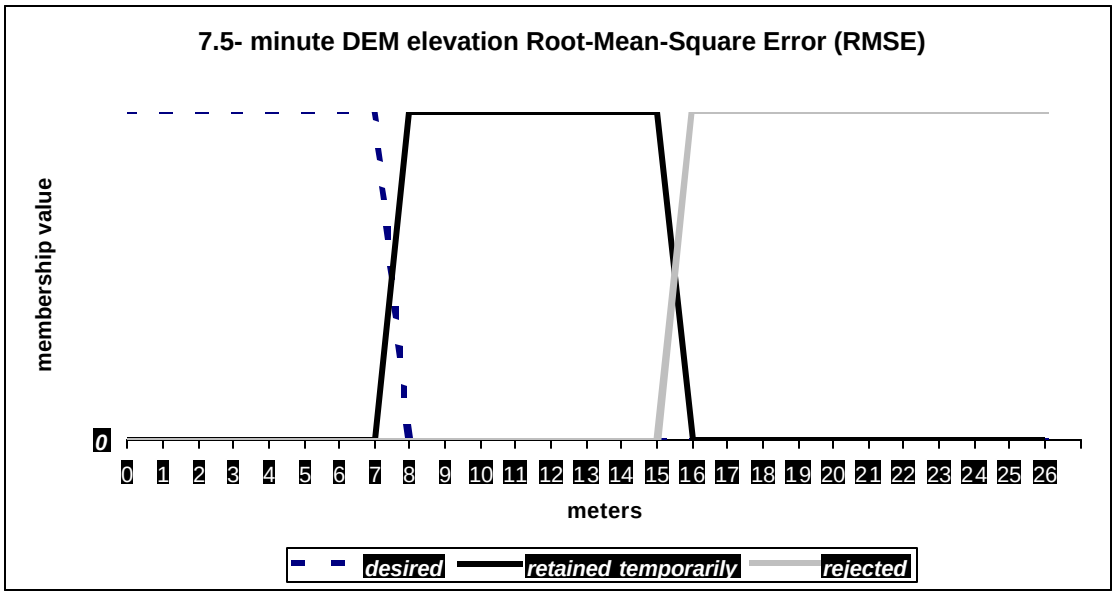
[4] Foley H., Petry F., Fuzzy Knowledge-Based System for Performing Conflation in Geographical Information Systems, 13th International Conference on Industrial &

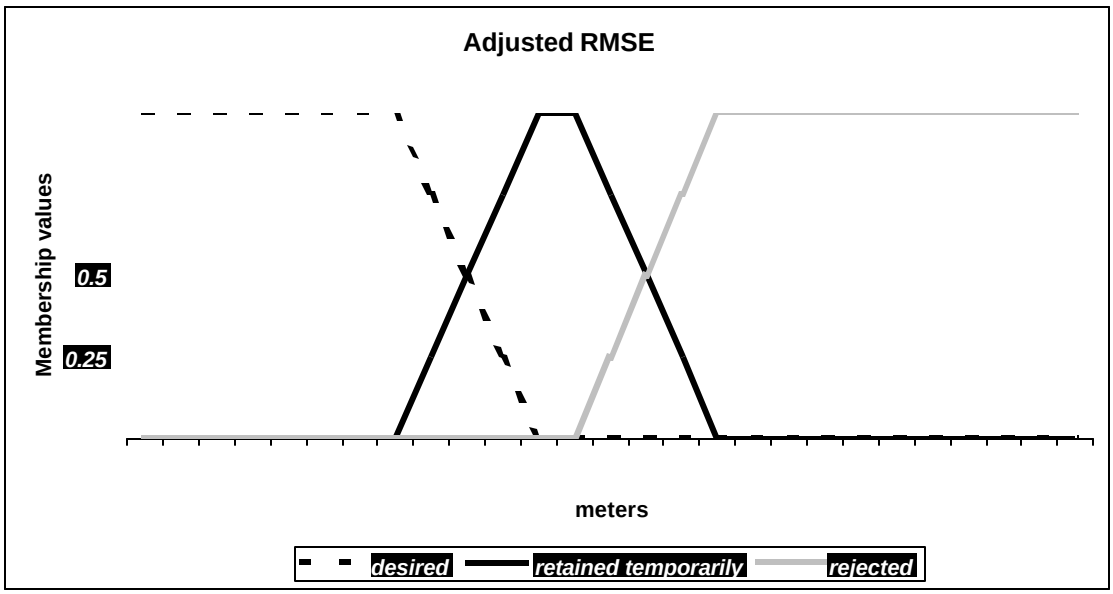
Engineering Applications of Artificial Intelligence & Expert Systems (EA/AIE-2000), June 19-22, 2000, New Orleans, http://www.cacs.usl.edu/~ieaaie2000/program.htm

[5] Jensen J., A. Saalfeld, F. Broome, K. Price, D. Ramsey, and L. Lapine, Spatial Data Acquisition and Integration, 2000, from NSF Workshop GIS and Geospatial Activities, http://www.ucgis.org/research_white/data.html

[6] Mark, D., Geographic information science: critical issues in an emerging cross-disciplinary research domain, 1999, NSF workshop to assess the needs for basic research in Geographic Information Science and Technology, January 14-15, 1999.

[7] Montello, D., Goodchild, M, Fohl, and J. Gottsegen (1998) Fuzzy spatial queries in digital spatial data libraries. Proceedings, FUZZ-IEEE 98, 1998 World, Congress on Computational Intelligence, Anchorage, Alaska.





Figures 3a and 3b. Illustrating the use of fuzzy logic membership functions