Formalism for representing data quality in non redundant spatial information

Viet PHANLUONG [(1)], Trung T. PHAM [(2)], Robert JEANSOULIN [(3)]

[(1)] Laboratoire d'Informatique Fondamentale de Marseille
[(3)] Laboratoire des Sciences de l'Information et des Systèmes
CMI, Université de Provence, 39 rue F. Joliot Curie, 13453 Marseille, France
Email: {phan, jeansoulin}@cmi.univ-mrs.fr

[(2)] CERMA – UMR CNRS 1563 Ambiances architecturales et urbaines
EAN – rue Massenet – BP81931 – 44319 Nantes Cedex 3, France
Email: trung.pham@cerma.archi.fr

ABSTRACT

The geographical information sources, resulting from a collection of different imperfect observations, often contain data redundancy. In this case we need to reduce the redundancy to obtain a non-redundant source before using it. Moreover, a geographical information source may contain data quality information, what means that the data source associates to any piece of information (geographically referenced), a value in the domain of one or several quality components. Geographical information pieces can sometimes overlap hence there is an issue on how to characterize the quality of the information within the intersection.

In this paper we consider the case, frequent with geographical information, where the semantic information domain is structured in a lattice, which is a particular case of hierarchy. Therefore, spatial overlapping can lead to redundancy or to conflicts. We present a formal background to reduce the redundancy and to answer questions about the assessment of quality when concurrent spatial information is combined within a common lattice structure. An algorithm to detect and reduce the redundancy, and to compute the resulting quality, is shown with an application example.

Keywords: Geographical information, data quality, lattice structure, information source.

1. Introduction

Geographical information can be defined as information about features and phenomena located on or near the terrestrial surface [6, 16]. Moreover, geographical data are created by abstracting phenomena existing in the real world. Hence, geographical information can be decomposed into primitive elements that take the form of (X, I) where: X refers to a general definition of a location of an object, and I stands for a list of properties or attributes of the phenomena observed at that location. I is also called semantic or descriptive information [16]. This 'relational' definition is more general than a 'functional' definition such as I = f(X), and is more appropriate for representing uncertain information.

When we observe a phenomenon, the result is restricted often on a set of (X, I), called an information source (data collection, database, knowledge base, etc.). An information source is defined as an identifiable collection of relative information. An information source can contain data quality information, if some, or all of the possible subsets of information, can be compared with respect to one or several quality elements (e.g.: from ISO19113). There are often many sources representing the same phenomena [7, 8].

The geographical information has the peculiar property to be referenced on a unique geographical space. Each reference is a two dimensional spatial extent (in our purpose). In general, different references may overlap: how to determine the information for the intersection?, and its quality? Moreover, a source can have different equivalent representations: how to evaluate the quality of information of these equivalent sources?

We start with the definition and formalization of a 'geospatial source' in the context where it exists some partially ordered structure in the I domain. We show that redundancy may exist, and we exhibit an algorithm for reducing it. Then we propose a representation for the data quality information, which behaves consistently between reduced and non-reduced versions of equivalent sources. The last section illustrates our approach on a 'land-cover' example: from the initial dataset, we compute a reduced equivalent version, and its quality.

2. Geospatial source

For manipulating and retrieving data, it is necessary to 'name' the information, to describe its 'meaning', and possibly to assess it for some particular purpose. These 'meanings' are represented often as a hierarchy [15] of categories grouped into classes, enabling the phenomena to be referred to at different levels of abstraction. This provides an essential means of distinguishing between information. Some data producers devise their own classification system for the phenomena of their own interest. For example the land-cover classification scheme from the CORINE project, presents an emphasis on natural and semi-natural vegetation [8].

The term 'hierarchy' is often used indiscriminately for any partial ordering [12, 14]. A lattice is a well suited structure for representing geographical information hierarchies, for classification, composition, aggregation, generalization, and abstraction [11]. A lattice can be modelled as a partially ordered set ($\mathcal{I}$, ≤) with two operators *meet* and *join* [3]. Here, we recall the notation of lattice, and we introduce the notation of 'geospatial source' as a set of couples (X, I), a relationship between the object set and the elements of an information lattice [9, 10].

2.1 Preliminaries: lattice and information lattice

An *ordered set* is a pair (I, ≤), where I is a set, and ≤ is a *partial order* on I:

$\forall a, b, c \in I$, $a \leq a$ (*reflexive*), $a \leq b$ and $b \leq a \rightarrow a = b$ (*anti-symmetric*), $a \leq b$ and $b \leq c \rightarrow a \leq c$ (*transitive*).

Let $A \subseteq I$, $A \neq \emptyset$. An element $a \in A$ is called a *minimal* (resp. *maximal}*) of A, if there exists no $x \in A$ such that $x \leq a$ (resp. $a \leq x$). The set of all minimal (resp. maximal) elements of A is noted *min*(A) (resp. *max*(A)). If a is the unique minimal (resp. maximal) element of A, then a is called the *least* (resp. *greatest*) element of A. An element $u \in I$ is an *upper bound* of A if for all $x \in A$, $x \leq u$. An element $l \in I$ is an *lower bound* of A if for all $x \in A$, $l \leq x$. Let u(A) denote the set of all upper bounds of A, and $l(A)$ denote the set of all lower bounds of A. The *least upper bound* of A, if exists, is denoted by $\vee A$, and called the *join* of A. The *greatest lower bound* of A, if exists, is denoted by $\wedge A$, and called the *meet* of A. In particular, if A = {a, b} then $\vee X$ is denoted by $a \vee b$, and $\wedge A$ is denoted by $a \wedge b$. We have the equivalence: $a \leq b$ iff $a = a \wedge b$ iff $b = a \vee b$. The least upper bound and the greatest lower bound of S, if exist, are denoted by $\top$ and $\bot$, respectively.

An ordered set $(I, \leq)$ is a $\vee$-*semi* (resp. $\wedge$-*semi*) *lattice* if $\forall a,b \in I$, $a \vee b$ (resp. $a \wedge b$) always exists. $(S, \leq)$ is a *lattice* if it is a $\vee$-semi lattice and a $\wedge$-semi lattice. It is a *complete lattice* if $\forall A \subseteq I$, $\vee A$ and $\wedge A$ always exist.

Definition 1 (Information lattice) An information lattice is a lattice which contains $\bot$, which represents the *inconsistency*, and $\top$, which represents the most general information.

Let a, $b \in \mathcal{I}$, if $a \leq b$, then b is '*more general*' than a (a is *more specific*). If $a \wedge b \neq \bot$, then a and b are '*complementary*'. If $a \wedge b = \bot$, then a and b are '*in conflict*'. The conflict is total if $a \vee b = \top$, as a and b do not share any common information. Otherwise, the conflict is partial, and $a \vee b$ is called the *agreement* of a and b.

In what follows, we consider an information lattice $(\mathcal{I}, \leq)$, and A, B being subsets of $\mathcal{I}$.

Definition 2 (Information containment) B is information contained in A, denoted by $A \angle B$, if for each $b \in B$ there exists $a \in A$ such that $a \leq$ b.

Intuitively, $A \angle B$, means B contains as much information as A. The relation $\angle$ is reflexive and transitive, but not anti-symmetric.

Definition 3 (Information equivalence) A is information equivalent to B, denoted by $A \approx B$, if $A \angle B$ and $B \angle A$.

Clearly, the relation $\approx$ is reflexive, symmetric, and transitive. We have some remarks: For any $A \subseteq \mathcal{I}$, we have: $A \approx min(A)$. For any A, $B \subseteq \mathcal{I}$, $A \approx B$ if and only if $min(A) \approx min(B)$.

Example 1:

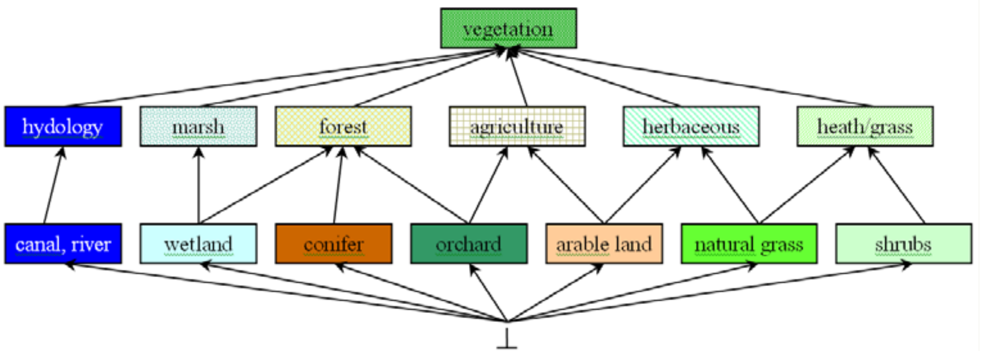


Figure 1: a part of information lattice of land cover classes

The graph in Figure 1 represents a hierarchy of land cover classes.

An arrow from a class a to a class b means that: $a \leq b$ or class a is less general (more specific) than class b, and there is no class c such that: $a \leq b \leq c$. On this information lattice, 'conifer' is more specific than 'forest', and we have min{'conifer', 'forest'} = {'conifer'}. {'orchard','forest'} is equivalent to {'orchard','agriculture'}. Consider now the sets of information {‘natural grass’} and {‘herbaceous’, ‘heath/grass’}, we have {‘herbaceous’, ‘heath/grass’} $\angle$ {‘orchard’}.

2.2 Sources and source mapping

To represent an observation, we need an object set $\mathcal{S}$ (the ‘*space*’). The observations can be incomplete, imprecise or ambiguous. Hence, it is not always appropriate to represent an observation directly by a function from $\mathcal{S}$ into $\mathcal{I}$. Rather, we adopt a relational representation: an observation is a finite list of (X, I), where $X \subseteq \mathcal{S}$ and $I \subseteq \mathcal{I}$. Indeed, in the registration process, the objects of $\mathcal{S}$ are given first, and the observations are picked in the power set of $\mathcal{I}$: $\wp(\mathcal{I})$, to build a set of (X, I).

Let’s give first some definitions: a *collection* of S is a set $X_1, X_{2,}..., X_k$ such that $X_i \subseteq S$, $1 \leq i \leq k$. A *covering* of X is a collection of S such that $S = \cup_{i=1,k} X_i$. A *partition* of S is a covering of S such that $i \neq j \rightarrow X_i \cap X_j = \varnothing$.

Definition 4 (geospatial source) Let $(\mathcal{I}, \leq)$ be an information lattice, and $\mathcal{S}$ be a non-empty space. An *information source* is a triple $\mathcal{D} = (P(\mathcal{S}), C(\mathcal{I}), R)$, where $P(\mathcal{S})$ is a covering of $\mathcal{S}$, $C(\mathcal{I})$ is a collection of $\mathcal{I}$, and R is a binary relation between $P(\mathcal{S})$ and $C(\mathcal{I})$, such that for each $X \in P(\mathcal{S})$ there exists $I \in C(\mathcal{I})$, such that $(X, I) \in R$.

The relation R is a set of pairs (X_i, I_i), where $X_i \in P(\mathcal{S})$ there exists $I_i \in C(\mathcal{I})$. By extension, we consider R represents $\mathcal{D}$. Moreover, for each $x \in X_i$, we consider x has all information in I_i.

Definition 4 (source mapping) The *source mapping* of a geospatial source $\mathcal{D} = (P(\mathcal{S}), C(\mathcal{I}), R)$ is a function f from $\mathcal{S}$ into $\wp(\mathcal{I})$ such that, for each $x \in \mathcal{S}$, $f(x) = \{info \in \mathcal{I} \mid \exists\ (X, I) \in R, x \in X, info \in I\}$.

On $\mathcal{S}$ we define the relation $\sim$ as follows: for any $x, y \in \mathcal{S}$, $x \sim y$ if and only if $f(x) \approx f(y)$. Clearly, $\sim$ is an equivalent relation. Let $X_1, X_{2,}..., X_k$ be the partition of $\mathcal{S}$ by the equivalent relation $\sim$. Thereby, the information source $\mathcal{D}$ can be restructured into the set of pairs (X_i, I_i), where $1 \leq i \leq k$, and $I_i = f(x)$, for every $x \in X_i$.

2.3 Equivalence of geospatial sources

With a relational representation as above, it is not immediate to distinguish between equivalent sources.

Definition 5 (Source equivalence) Let $\mathcal{D}_1$ and $\mathcal{D}_2$ be two geospatial sources on a same space $\mathcal{S}$ and a same information lattice $(\mathcal{I}, \leq)$. Let f_1 and f_2 be the respective source mappings. $\mathcal{D}_1$ is called $\approx$-equivalent to $\mathcal{D}_2$, denoted by $\mathcal{D}_1 \cong \mathcal{D}_2$, if for each $x \in \mathcal{S}$, $f_1(x) \approx f_2(x)$.

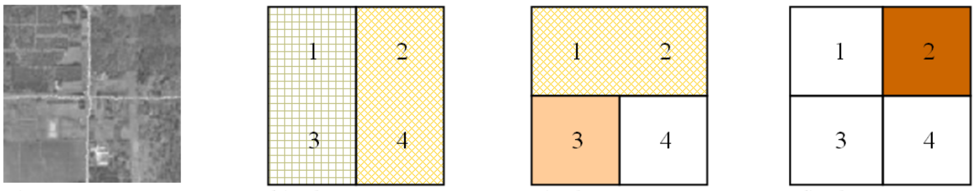
Geospatial sources can be redundant in sense of $\approx$-equivalence. Consider a pair (X, I) in R, if $min(I)$ is strictly included in I, then the pair (X, I) has *information redundancy*, since $min(I) \approx I$. Consider now (X_1, I_1) and (X_2, I_2) in R, if $I_2 \angle I_1$ and $X_1 \cap X_2 \neq \varnothing$, then pair $(X_2,$

I_2) contains *object redundancy*, as the information associated with $X_1 \cap X_2$ by (X_2, I_2) can be deduced from (X_1, I_1), in sense of information containment.

A geospatial source $\mathcal{D}$ is said *reduced* if it has neither information redundancy, nor object redundancy. A *reduced* geospatial source $\mathcal{D}'$ is ≅-equivalent to $\mathcal{D}$. The *Algorithm Reduce* computes the reduction of $\mathcal{D}$.

```
Algorithm Reduce
Input: an information source $\mathcal{D} = (P(\mathcal{S}), C(\mathcal{I}), R)$
Output: $\mathcal{D}' = (P'(\mathcal{S}), C'(\mathcal{I}), R')$ reduced and equivalent to $\mathcal{D}$.
Procedure:
        1. (Mapping) Computing the source mapping $f$ of $\mathcal{D}$ and result in $R_f$.
        2. (Redundancy reduction)
                For each pair $(x_i, I_i) \in R$, where $x_i$ is an element of $\mathcal{S}$ and $I_i = f(x_i)$
                        Replace $(x_i, I_i)$ by $(x_i, min(I_i))$.
        3. Return the result of step 2.
```

***Example* 2**: we have different observations on a space $\mathcal{S}$, which is partitioned in 4 blocs p_1, p_2, p_3, p_4, and we use the elements of the information lattice ($\mathcal{I}$, ≤) as in Figure 1 (see Example 1) to describe the result of observation.



a) *a region* b) *observation 1* c) *observation 2* d) *observation 3*

Figure 2: example of space partition and land cover classes

Suppose that in some sources, some blocs (e.g: p_1, p_2) are not independently observed, as in the Table 1: p_2 is associated twice to 'forest', hence the source is object-redundant. The source mapping f of $\mathcal{D}$ is in Table 2.

Table 1: Information source $\mathcal{D}$

Objects	Information
p_1, p_3	agriculture
p_2, p_4	forest
p_3	arable land
p_1, p_2	forest
p_2	conifer

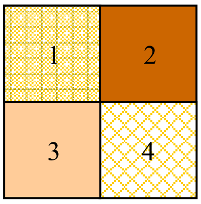
Table 2: Source mapping

Objects	Information
p_1	forest, agriculture
p_2	conifer, forest
p_3	arable land, agriculture
p_4	forest

In source mapping, p_2 is associated to {'conifer', 'forest'}. There exists an information redundancy because 'conifer' is a specific of 'forest'. The result of step 2 is in Table 3.

Table 3: Reduction of information source $\mathcal{D}$

Objects	Information
p_1	forest, agriculture
p_2	conifer
p_3	arable land
p_4	forest



Reduction of $\mathcal{D}$

These sources are equivalent, but the last table is a compact representation which not contains the object redundancy and the information redundancy.

3. Evaluation basis of the equivalent source

3.1 Data quality and notion of quality domain

Geographical information is related to a universe (real world). The *universe* is always represented according to a modelling of the information which corresponds to the needs of the data producer, with the knowledge that he has of this universe and with the sources of existing information. This modelling determines the specification of the geographical database. The specification of data defines a particular view of the real world. To be able for defining the parameters of data quality, it is necessary to model in a precise way the physical universe to lead to the dataset. The universe filtered by the specification is called the *nominal terrain*, which is the virtual image of the universe at a given date. The nominal terrain finds all its interest when one treats the *internal quality* which can be reformulated like the adequacy between the nominal terrain and the really produced dataset [5].

The *external quality* is relative to the needs of users. It presents the fitness of a dataset for a potential application. There are different standards of data quality as SDTS, ICA, CEN/TC287, ISO/TC211. The principal aspects of data quality are *lineage*, *accuracy* (spatial accuracy, thematic accuracy, and temporal accuracy), *precision*, *consistency*, *completeness*, *timeliness* [2, 7, 13]. In this part, we discuss on the aspects of data quality, in the context of geographical information [1, 7].

Here, we consider quality aspects that can be modelled by total ordered sets as lineage, accuracy, and completeness. For dataset acquired from aerial photography, date of photography, coordinate system and map projection are used as the information of lineage. The user must be able to assess the dataset from the point of his particular application. The user can refer the source most recent or the source close to the point of his application. For the accuracy aspect, positional accuracy expresses the degree of discrepancy between the encoded location and the location defined in the specification. For lines and areas the situation is more complex. Positional accuracy for lines and areas are measured by the positional accuracy of points that define those lines. In general each quality aspect may have one or more “metrics” which is quantitative statement of the quality aspect. Metrics may be statistical (expressed in real numbers, with or without units) or binary and in some instances may be descriptions. For the completeness, errors of omission, errors of commission and thematic accuracy can be expressed by percentages or probabilities.

We call a *quality domain* associated with an information source $\mathcal{D}$ a total ordered set of values that defines a quality attribute of information in $\mathcal{D}$. A quality domain can be a set of numeric values (e.g, percentages, real numbers, binary numbers, etc.) or symbolic values (e.g., slow, medium, fast; bad, medium, good, etc.) which estimate a quality aspect of information. In general an information source can be associated with one or many quality domains as lineage, geometric accuracy, thematic accuracy, errors of commission, errors of omission, or up-to-date, or a composition of the quality components, etc. In what follows, the order of a quality domain is denoted by $\leq$.

We consider an information piece (X, I) such as a dataset and its quality is estimated in data collection. Thus, a source containing quality information on a quality component Q is defined as a collection of information pieces (X, I) such that each information piece is associated to a value of Q. In other words, a source containing quality information is a set of triple $(X, I, qoXI)$ where $qoXI$ is quality information on a given quality domain. In the next, $qoXI$ is denoted by $q(X, I)$.

3.2 Data quality formalism

An information source can have different, but equivalent representations. We suppose that a quality information is attributed to the original source. How can we attribute quality to the source obtained by the reduction of the original source? In what follows, we provide a formal basis to answer this question.

Definition 7 (*Q*-quality) Let $\mathcal{D} = (P(\mathcal{S}), C(\mathcal{I}), R)$ be an information source. Let Q be a quality domain associated with $\mathcal{D}$. A function q from R into Q, is called Q-*quality* (quality on domain Q of each information piece) of source $\mathcal{D}$, such that for any $(X_1, I_1), (X_2, I_2) \in R$, the following conditions are satisfied:

a) Let $X_1 \subseteq X_2$ and $I_2 \angle I_1$, then $q(X_1, I_1) \geq q(X_2, I_2)$.
b) If $I_1 = \{ \top \}$, then $q(X_1, I_1) = max(Q)$.

This condition specifies that a quality function is anti-monotonic with respect to object set and information set. We remind that after Definition 4, the notion of source mapping implies that for each pair $(X, I) \in R$, for each $x \in X$, the source $\mathcal{D}$ associates x with all information elements in I. As a consequence, if (X_2, I_2) is in R, and for any $X_1 \subseteq X_2$ and for any $I_2 \angle I_1$, the pair (X_1, I_1) is deduced from (X_2, I_2). Therefore, if (X_2, I_2) is viewed with quality $\mathrm{q}(X_2, I_2)$, then (X_1, I_1) is viewed with at least the quality $\mathrm{q}(X_1, I_1)$.

Moreover, with respect to geographic information, we remark that Definition 7 is scale-independent (not concerned with resolution). In fact, both X_1 and X_2 are contained in $\mathcal{S}$, which is obtained with a fixed spatial resolution. Hence, the condition $X_1 \subseteq X_2$ does not mean X_1 corresponds to a finer resolution than X_2. A direct consequence of Definition 5 is: if $X_1 = X_2$ and $I_1 \approx I_2$ then $q(X_1, I_1) = q(X_2, I_2)$. Further, if $\mathcal{D}_1$ and $\mathcal{D}_2$ are equivalent sources, with respective source mapping f_1 and f_2, then for any $x \in \mathcal{S}$, $q(\{x\}, f_1(x)) = q(\{x\}, f_2(x))$.

3.3 Evaluation of the quality of equivalent sources

Let $\mathcal{D} = (P(\mathcal{S}), C(\mathcal{I}), R)$ be an information source, and q a Q-quality for $\mathcal{D}$. Let $\mathcal{D}' = (P'(\mathcal{S}), C'(\mathcal{I}), R')$, be an equivalent reduced source of $\mathcal{D}$. How can we attribute quality for $\mathcal{D}'$? To answer this question, we propose a formalism to evaluate approximately the quality of the equivalent sources.

Definition 8 (extension of *Q*-quality) Let q be a Q-quality of $\mathcal{D} = (P(\mathcal{S}), C(\mathcal{I}), R)$. A Q-quality extension of q is a Q-quality of source $\mathcal{D}' = (P'(\mathcal{S}), C'(\mathcal{I}), R')$ which is equivalent to $\mathcal{D}$, denoted by q' such that for any $(X_k, I_k) \in R'$ and $(X_k, I_k) \in R$, $q'(X_k, I_k) = q(X_k, I_k)$.

Definition 8 requires that q' must be a correct extension of q. In what follows, we show the existence of an extension of a Q-quality. We build a function $q'(X, I)$, an extension of q. Consider a pair $(X, I) \in R'$. Since $P(\mathcal{S})$ is a covering of $\mathcal{S}$, and $X \subseteq \mathcal{S}$ there exists $X_i \in P(\mathcal{S})$ such that $X \cap X_i \neq \varnothing$ and the union of all such $X \cap X_i$ is equal to X. Let q' be the function defined as follows:

Case 1: If $I = \{ \top \}$, then define $q'(X, I) = max(Q)$.
Case 2: Else, if there exists $(X_i, I_i) \in R$ such that $X \subseteq X_i$ and $I_i \angle I$, then define
$$q'(X, I) = max\{ q(X_i, I_i) \mid (X_i, I_i) \in R, X \subseteq X_i \text{ and } I_i \angle I \}$$
Case 3: Else, let $K = \wedge\{I_i \mid (X_i, I_i) \in R, X \cap X_i \neq \varnothing \}$.
Case 3.1: If $K \angle I$, then define $q'(X, I) = min\{ m(X_i, I_i) \mid (X_i, I_i) \in R, X \cap X_i \neq \varnothing \}$,
where $m(X_i, I_i) = max\ \{q(X_j, I_j) \mid (X_j, I_j) \in R, X_i \cap X \subseteq X_j \cap X \}$.
Case 3.2: Else, define $q'(X, I) = min(Q)$.

We can show from the definition, that q' is a well-defined function, and for any case if $X_1 \subseteq X_2$ and $I_2 \angle I_1$, then $q'(X_1, I_1) \geq q'(X_2, I_2)$. Hence, q' is a correct Q-quality extension of q.

Using this extension, we can define the Q-quality for any equivalent representation of $\mathcal{D}$, in particular, the reduction of $\mathcal{D}$.

Example 3: Let $\mathcal{D} = (P(\mathcal{S}), C(\mathcal{I}), R)$ the information source of Table 1 of Example 2. Consider the 'semantic accuracy' as a quality taking values in the domain: {*h*-high, *m*-medium, *l*-low, *p*-poor} such as $h > m > l > p$. The semantic accuracy values, for the source $\mathcal{D}$, are provided by Table 4. This semantic accuracy satisfies the conditions of Q-quality in the Definition 7.

Consider the information pieces, that we may also name 'formal concepts' of the lattice: ({ p_3},{'arable land'}) and ({ p_1, p_3},{'agriculture'}), we see that: { p_3} $\subseteq$ { p_1, p_3} with {'agriculture'} $\angle$ {'arable land'}, so this satisfies the conditions of Q-quality.

The reduction of $\mathcal{D} = (P(\mathcal{S}), C(\mathcal{I}), R)$ is $\mathcal{D}' = (P'(\mathcal{S}), C'(\mathcal{I}), R')$, given in Table 3. For q' column in Table 5, the semantic accuracy values for the concepts ({ p_2},{'conifer}) and ({ p_3}, {'arable land}) are given directly from Table 4 by Definition 8, but for ({ p_1},{'forest','agriculture'}) and ({ p_4},{'forest'}) we need to give a value consistent with the general Q-quality Definition 7.

- For ({ p_1},{'forest','agriculture'}), the Case 3.1 applies:

$$\{(X_i, I_i) \in R, X \cap X_i \neq \varnothing \} = \{(\{ p_1, p_3\},\{\text{'agriculture'}\}), (\{p_1, p_2\}, \{\text{'forest'}\})\}.$$

K= $\wedge${'agriculture','forest'}= {'agriculture'}$\wedge${'forest'}= {'orchard'}, and {'orchard'}$\angle${'forest','agriculture'}.

$q'(X, I) = min\{ q(X_i, I_i) \mid (X_i, I_i) \in R, X \cap X_i \neq \emptyset \} = min\{q(\{ p_1, p_3\},\{\text{‘agriculture’}\}), q'(\{p_1, p_2\},\{\text{‘forest’}\})\} = l$. It means that it is the confusion between ‘agriculture’ and ‘forest’ in the parcel p_1.

- For $(\{ p_4\},\{\text{‘forest’}\})$, the Case 2 applies, because there exist unique $(X_i, I_i) = (\{ p_2, p_4\},\{\text{‘forest’}\})$ such that $X \subseteq X_i$ and $I_i \angle I$. Therefore $q'(\{ p_4\},\{\text{‘forest’}\}) = q(\{ p_2, p_4\},\{\text{‘forest’}\}) = m$.

Table 4: Information source $\mathcal{D}$

Objects	Information	Semantic accuracy (q)
p_1, p_3	agriculture	l
p_2, p_4	forest	m
p_3	arable land	p
p_1, p_2	forest	m
p_2	conifer	h

Table 5: Reduced source

Objects	Information	Semantic accuracy (q’)
p_1	forest, agriculture	l
p_2	conifer	h
p_3	arable land	p
p_4	forest	m

4. Application example

We provide an example to illustrate our approach, using two observations of a region (North Carolina example) with hypothetical data quality (not provided on the web server). The information lattice of the land cover classes in the Figure 1, is used to observe the region, and two observations, and their quality, are in the Figure 3 and 4. The domain of semantic accuracy is $\{h, m, l, p\}$ with $h > m > l > p$ as above.

The information source is the collection of two observations with the object set as the result of the complete spatial intersection. It is impossible visualize this original source in a cartographic form, because it contains the redundancies. So we apply our approach to reduce this source and evaluate the quality of the result. Finally, the figure 5 shows the result.

5. Conclusion

We have proposed an approach to perform the reduction of a geographical information source to a non-redundant equivalent source, in the case where the semantic information has a lattice structure. When the source results from imperfect observations (e.g.: a space coverage which is not a partition, an imperfect survey), and if some information about the quality of the different information pieces is available, we need to estimate the quality for the equivalent reduced source. Our approach has been experimented on some sample land-cover data, and we can conclude that they behave correctly, meaning that the resulting quality is consistent with the original data.

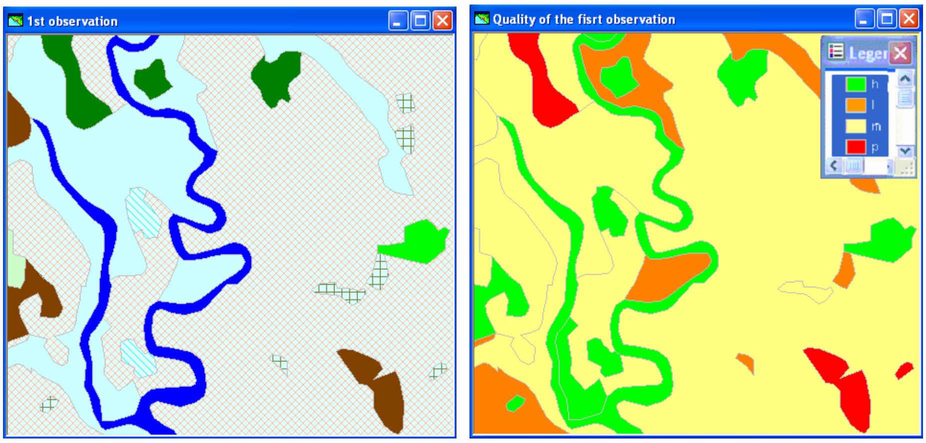


Figure 3: first observation and quality

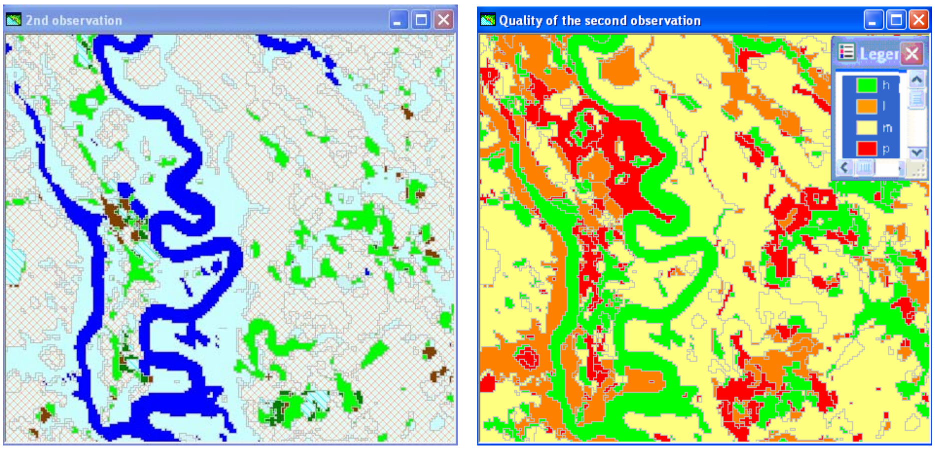


Figure 4: second observation and quality

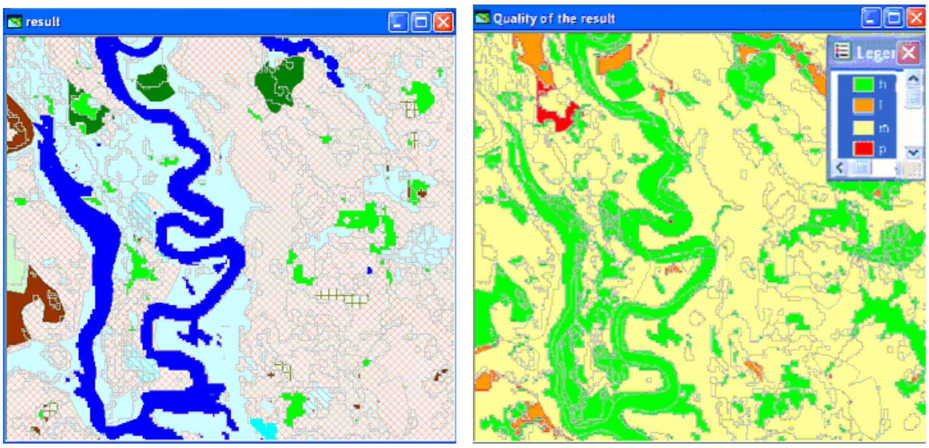


Figure 5: Result of spatial information reduction and its quality

References

[1] H. Aalders. The registration of quality in a GIS. In *the proceeding of the first International Symposium on Spatial Data Quality*, pages 23-32, Hongkong, 1999.

[2] H. Aalders and J. Morrison. Spatial data quality for GIS. In Massimo Craglia and Harlan Onsrud, editors, *Geographic Information Research: Trans Atlantic Perspectives*, pages 463-475. Taylor & Francis, 1998.

[3] S. Burris and H.P. Sankappanavar. *A course in universal Algebra.* Springer-Varlag, 1981.

[4] N.R. Chrisman. The role of quality in the long term functioning of geographical information system. In *Proceeding of international symposium of automated cartography,* pages 303-321, Ottawa, Canada, 1983.

[5] B. David and P. Fasquel. Qualité d'une base de données géographiques: concept et terminologie. Technical report 67, IGN, 1997.

[6] M.F. Goodchild, M.J. Egenhofer, K.K. Kemp, D.M. Mark, and E. Sheppard. Introduction to the varenius projet. In *international Journal of Geographical Information Science,* 0(13): 731-745, 1999.

[7] A. Jakobson. Quality Evaluation of Topographic Datasets - Experiences in European National Mapping Agencies. In *the proceeding of the 1st international symposium on spatial data quality,* pages 154-164, Hongkong, 1999.

[8] C. Jones. Geographical information concepts and spatial models. In *the Geographical Information Systems and Computer Cartography*, pages 18-38, Longman Publishers Ltd, 1997.

[9] V. PhanLuong, T.T. Pham, and R. Jeansoulin. Data quality based fusion: application to land cover. In *the proceeding of 7th international conference of information fusion (Fusion04), pages 672-679, Stockholm, Swenden, 2004.*

[10] V. PhanLuong, T.T. Pham, and R. Jeansoulin. Data integration under lattice structure. In *the proceeding of 14th international symposium on methodologies for intelligent systems (ISMIS'03), pages 83-88, Maebashi, Japan, 2003.*

[11] C.S. Smyth. *Spatial and Temporal Reasoning in Geographic Information Systems*. chapter A Representational Framework for Geographic Modelling, pages 1996, Oxford University Press, Inc, 1998.

[12] J. Sowa. Building, sharing and merging ontologies. 2001, URL: http:/www.jfsowa.com/ontology/ontoshar.htm

[13] H. Veregin. Data quality parameters. *Journal of Geographical Information Systems*. 1:177-189, 1999.

[14] M.F. Worboys. Computation with imprecise geospatial data. *Computers, Environment and Urban Systems*, 2(22):85-106; 1998.

[15] M.F. Worboys and M. Duckham. Integrating spatio-thematic information. In *the proceeding of international conference GIScience,* 2002.

[16] J. Zhang and M. Googchild. *Uncertainty in Geographical Information*, chapter Geographical Information and Uncertainty, pages 1-13. London: Taylor and Francis, 2002.