Data Quality Based Fusion: Application to Land Cover

V. Phan Luong
LIF, CNRS UMR6166
Université de Provence
39 rue F. Joliot Curie
13453, Marseille
France
phan@cmi.univ-mrs.fr

T.T. Pham
CNRS UMR6168
Université de Provence
39 rue F. Joliot Curie
13453, Marseille
France
pham@cmi.univ-mrs.fr

R. Jeansoulin
CNRS UMR6168
Université de Provence
39 rue F. Joliot Curie
13453, Marseille
France
jeansoulin@cmi.univ-mrs.fr

Abstract – *This paper proposes an approach to integrate redundant, complementary, and/or conflicting information sources, assuming that information space is a lattice, and metadata about data quality is available. The approach is then applied on land cover assessments, where land cover classes are in a hierarchical structure, and quality information is defined, based on completeness and consistency of attributed classes in the assessments. Experimental results show that the approach really improves the assessments.*

Keywords: data fusion, data quality, information lattice, land cover.

1 Introduction

Data fusion has been defined in [7] as a "formal framework in which are expressed means and tools for the alliance of data of the same scene, originating from different sources. It aims at obtaining information of greater quality; the exact definition of greater quality will depend upon the application.". Information from several sources can be heterogeneous and associated with various quality levels. This paper focuses on a framework to integrate redundant, complementary, conflict information, assuming that information has a lattice structure [4], and the quality of information is available in each source. Instances of this problem occur in merging ontologies [2, 5], integrating geographical information [8, 9]. In particular, classes of land cover problem in geographical information are in hierarchy structure, that can be mapped to a lattice structure.

The paper is organized as follows: Section 2 recalls the work in [4], which presents the formal basis for our approach. The notions of redundant, complementary or conflicting data are formalized, and the methods for integration under lattice structure are given. Section 3, we formalize the notion of quality, and extend the above methods in the context where quality information is available. Section 4 presents a case study of land cover assessment. We consider the basic information of the land cover LCM2000, and two different methods of assessment. We show how our methods of integration can be applied in this case to obtain a better assessment. In this purpose, we define a data quality measure which depends on the completeness and the consistency of data. Finally, concluding remark and future work is presented in section 5.

2 Source fusion under information lattice

2.1 Information lattice

Definition 1 - Information lattice. *An* information lattice *is a lattice $(\mathcal{I}, \preceq)$ which contains $\bot$ and $\top$.*

The element $\bot$ represents unknown. The element $\top$ represents inconsistent. For all $x \in I, \bot \preceq x$, and $x \preceq \top$. Let $x, y \in \mathcal{I}$. If $x \prec y$ then y is called more *complete* or more *specific* than x. Let $X \subseteq \mathcal{I}$ such that $X \neq \emptyset$. The set of all minimal (resp. maximal) elements of X is denoted by $min(X)$ (resp. $max(X)$). The least upper bound of X, if exists, is denoted by $\vee X$, and called the *join* of X. The greatest lower bound of X, if exists, is denoted by $\wedge X$, and called the *meet* of X. In particular, if $X = \{x, y\}$ then $\vee X$ and $\wedge X$, respectively denoted by $x \vee y$ and $x \wedge y$, always exist. We have $x \wedge y \preceq x, y \preceq x \vee y$, and the following equivalence: $x \preceq y$ iff $x = x \wedge y$ iff $y = x \vee y$.

If $x \vee y \neq \top$, then x and y are called *complementary* to one another. If $x \vee y = \top$ then x and y are called in *conflict* with each other. The conflict between x and y is called total if $x \wedge y = \bot$, as x and y do not share any common information. Otherwise, the conflict is called partial, and $x \wedge y$ is called a *consensus* of x and y.
In what follows, we consider an information lattice $(\mathcal{I}, \preceq)$.

Definition 2 - Information containment. *Let I and J be subsets of $\mathcal{I}$. We define $I \sqsubseteq J$ if $I = \emptyset$, or for each $x \in I$, there exists $y \in J$ such that $x \preceq y$.*

Intuitively, $I \sqsubseteq J$ means that J contains as much information as I. The relation $\sqsubseteq$ generalizes the set inclusion $\subseteq$. Indeed, if $I \subseteq J$ then $I \sqsubseteq J$ (by the reflexivity of $\preceq$). But the inverse is not true. Moreover, $\sqsubseteq$ is reflexive and transitive, but not antisymmetric. For instance, consider the ordered set $(N, \leq)$, where N is the set of natural numbers, and $\leq$ is the usual order on N. Let $I = \{3, 5, 6\}$ and $J = \{4, 6\}$. After Definition 2, $I \sqsubseteq J$ and $J \sqsubseteq I$, but $I \neq J$.

Definition 3 - Information equivalence. *Let I, J be subsets of $\mathcal{I}$. Define $I \simeq J$ iff $I \sqsubseteq J$ and $J \sqsubseteq I$.*

Clearly, the relation $\simeq$ is reflexive, symmetric, and transitive. For any $I, J \subseteq \mathcal{I}$, we have $I \simeq max(I)$, and $I \simeq J$ if and only if $max(I) = max(J)$.

Definition 4 - Consensus and Aggregation. *Let I and J be subsets of $\mathcal{I}$. If I and J are non-empty, then define the consensus and the aggregation of I and J to be respectively $I \otimes J = max(\{x \wedge y \mid x \in I, y \in J\})$, and $I \oplus J = max(J)$ (resp. $max(I)$), if $I \sqsubseteq J$ (resp. $J \sqsubseteq I$). Otherwise, $I \oplus J = max(\{x \vee y \mid x \in I, y \in J\})$. If $I = \emptyset$ (or $J = \emptyset$), then define $I \otimes J = \emptyset$, and $I \oplus J = max(J)$ (resp., $max(I)$).*

The consensus and aggregation operations have the following properties:

(p1) Let I_1 and I_2 be subsets of $\mathcal{I}$. If $I_1 \simeq I_2$ then for any $J \subseteq \mathcal{I}$, we have $I_1 \otimes J \simeq I_2 \otimes J$, and $I_1 \oplus J \simeq I_2 \oplus J$.

(p2) Let I and J be subsets of $\mathcal{I}$. We have:

(a) $I \otimes J \sqsubseteq I$, and $I \sqsubseteq I \oplus J$, and

(b) $I \cap J \sqsubseteq I \otimes J \sqsubseteq I \cup J \sqsubseteq I \oplus J$.

2.2 Information source

A (finite) *collection* of a set X is a set $\{X_1, ..., X_k\}$ such that $X_i \subseteq X, 1 \leq i \leq k$. A (finite) *covering* of a set X is a collection $\{X_1, ..., X_k\}$ of X such that $\cup_{i=1,k} X_i = X$.

Definition 5 - Information Source. *Let $\mathcal{I}$ be an information lattice, and $\mathcal{S}$ be a non-empty set of objects, called object 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 $C(\mathcal{I})$.*

We call the *source mapping* of $\mathcal{D}$ the function f from $\mathcal{S}$ into $2^{\mathcal{I}}$ such that for each $x \in S$,

$$f(x) = \{i \in \mathcal{I} \mid \exists (X, I) \in R, x \in X, i \in I\}$$

On $\mathcal{S}$ we define an equivalent relation $\sim$ as follows: for any $x, y \in \mathcal{S}, x \sim y$ if and only if $f(x) = f(y)$.

Different information sources can essentially represent the same information. In such a case, we consider they are equivalent. Precisely, we have the following definition:

Definition 6 - Source equivalence. *Let $\mathcal{D}_1$ and $\mathcal{D}_2$ be two information sources on a same object space $\mathcal{S}$ and a same information lattice $\mathcal{I}$. Let f_1 and f_2 be the source mappings of $\mathcal{D}_1$ and $\mathcal{D}_2$, respectively. $\mathcal{D}_1$ is called* $\simeq$-equivalent *to $\mathcal{D}_2$, denoted by $\mathcal{D}_1 \simeq \mathcal{D}_2$, if for each $p \in \mathcal{S}$, $f_1(p) \simeq f_2(p)$.*

Information sources can contain redundancies in sense of $\simeq$-equivalence. Consider pair (X, I) in R, if $max(I)$ is strictly included in I, then the pair (X, I) has *information redundancy*, since $max(I) \simeq I$. Consider now pairs (X_1, I_1) and (X_2, I_2) in R, if $I_1 \sqsubseteq I_2$ and $X_1 \cap X_2 \neq \emptyset$, then pair (X_1, I_1) contains *object redundancy*, as the information associated with $X_1 \cap X_2$ by (X_1, I_1) can be deduced from (X_2, I_2).

An information source $\mathcal{D}$ is called *reduced* if $\mathcal{D}$ has no information redundancy and no object redundancy. 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 $\simeq$-equivalent to $\mathcal{D}$.
Method:
1. Computing the source mapping f of $\mathcal{D}$.
2. Let $R_\sim$ be the set of pairs (X_i, I_i), where X_i are in the partition of $\mathcal{S}$ by $\sim$, and $I_i = f(p)$ for every $p \in X_i$.
3. Replace each pair (X, I) in $R_\sim$ by $(X, max(I))$.
4. Regrouping: In the result of step 3, replace all pairs (X_1, I_1), ..., (X_k, I_k) such that $I_1 = ... = I_k = I$ by the pair $(X_1 \cup ... \cup X_k, I)$.
5. Return the final result of step 4.

2.3 Integration of information sources

Let $\mathcal{D}_1$ and $\mathcal{D}_2$ be two information sources on a same object space $\mathcal{S}$ and a same information lattice $\mathcal{I}$. Let f_1 and f_2 be the source mappings of $\mathcal{D}_1$ and $\mathcal{D}_2$, respectively. An integration of $\mathcal{D}_1$ and $\mathcal{D}_2$ is an information source $\mathcal{D} = (P(\mathcal{S}), C(\mathcal{I}), R)$ such that the source mapping f of $\mathcal{D}$ satisfies the following: for each $p \in \mathcal{S}$, $f_1(p) \otimes f_2(p) \sqsubseteq f(p) \sqsubseteq f_1(p) \oplus f_2(p)$. We denote $f = f_1 \theta f_2$, and call $\mathcal{D}$ the θ-integration of $\mathcal{D}_1$ and $\mathcal{D}_2$.

- If for all $p \in \mathcal{S}, f(p) = f_1(p) \theta f_2(p) = f_1(p) \otimes f_2(p)$, then f is called the *pessimistic integration*.

- If for all $p \in \mathcal{S}, f(p) = f_1(p) \theta f_2(p) = f_1(p) \oplus f_2(p)$, then f is called the *optimistic integration*.

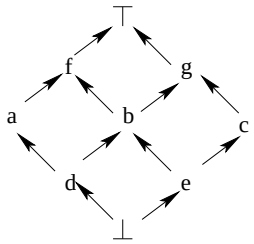
Algorithm θ-Integrate integrates $\mathcal{D}_1$ and $\mathcal{D}_2$:

ALGORITHM θ-Integrate
Input: $\mathcal{D}_1 = (P_1(\mathcal{S}), C_1(\mathcal{I}), R_1)$ and $\mathcal{D}_2 = (P_2(\mathcal{S}), C_2(\mathcal{I}), R_2)$.
Output: $\mathcal{D}$ an integration of $\mathcal{D}_1$ and $\mathcal{D}_2$.
Method:
1. Let $\mathcal{D}_1' =$ Reduce($\mathcal{D}_1$) and $\mathcal{D}_2' =$ Reduce($\mathcal{D}_2$).
2. Let $\mathcal{D} = (P(\mathcal{S}), C(\mathcal{I}), R)$, where $P(\mathcal{S}) = \emptyset$, $C(\mathcal{I}) = \emptyset$, and $R = \emptyset$.
3. For each pair (X_1, I_1) in $\mathcal{D}_1'$ do
   For each pair (X_2, I_2) in $\mathcal{D}_2'$ do begin
   Produce $(X_1 \cap X_2, I_1 \theta I_2)$;
   If $X_1 \cap X_2 \neq \emptyset$, then insert $X_1 \cap X_2$ into $P(\mathcal{S})$, $I_1 \theta I_2$ into $C(\mathcal{I})$, and $(X_1 \cap X_2, I_1 \theta I_2)$ into R;
   end;
4. Return Reduce($\mathcal{D}$).

EXAMPLE 1 *Let $\mathcal{S} = \{1, 2, 3, 4\}$, $\mathcal{I} = \{a, b, c, d, e, f, g, \perp, \top\}$ with the partial order $\preceq$ defined by the Hasse diagram in Figure 1. Let $\mathcal{D}_1$ and $\mathcal{D}_2$ be information sources on the object space $\mathcal{S}$ and the information lattice $\mathcal{I}$ as in Figure 1 (a) and (b). The reductions of sources following algorithm Reduce are in Figure 1 (c) and (d). The results of integrations are in Figure 1 (e), (f), (g), and (h).*

3 The quality based approach

Data quality, which is part of metadata associated with data in an information source, is relative to the needs of users. A commonly used definition of quality is 'fitness for use'.



(a)

R_1	
P_1(S)	C_1(I)
1	a, d, e, f
2	a, b, e
3	b, d, e
1,4	a, f
2,3	b, e

(b)

R_2	
P_2(S)	C_2(I)
1	b, c
2	a, e
3	a, b
1,3	b
3,4	d, e

(c)

Reduce(R_1)	
P_1'(S)	C_1'(I)
1, 4	f
2	a, b
3	b

(d)

Reduce(R_2)	
P_2'(S)	C_2'(I)
1	b, c
2	a, e
3	a, b
4	d, e

(e)

θ −Integration	
P_θ(S)	C_θ(I)
1	{b, c} θ {f}
2	{a, b} θ {a, e}
3	{b} θ {a, b}
4	{f} θ {d, e}

(f)

Pessimistic integ.	
P_{pes}(S)	C_{pes}(I)
1, 3	b
2	a, e
4	d, e

(g)

Optimistic integ.	
P_{opt}(S)	C_{opt}(I)
1	⊤
2, 3	a, b
4	f

(h)

Intermediate integ.	
P_{int} (S)	C_{int}(I)
1	c, f
2, 3	a, b
4	f

Fig. 1: An example of information lattice, reduction and integration of sources

There are many aspects of data quality [6] as correctness, accuracy, precision, completeness, consistency, relevance, and timeliness, etc. In this work, we consider quality aspects that can be modeled by total ordered sets. In effect, we define a *quality domain* associated with an information source $\mathcal{D}$ to be a *totally ordered* set of values. A quality domain can be a set of symbolic values or numeric values. In general an information source can be associated with one or many quality domains as precision, certitude, completeness, etc. In what follows, the order of a quality domain is denoted by $\leq$.

3.1 The quality basis

Definition 7 *Let $\mathcal{S}$ and $\mathcal{I}$ be respectively an object space and an information space. Let Q be a quality domain. A Q-*quality *on $\mathcal{S} \times \mathcal{I}$ is a function ϕ from $(2^{\mathcal{S}} \setminus \{\emptyset\}) \times (2^{\mathcal{I}} \setminus \{\emptyset\})$ into Q, such that for any $(X_1, I_1), (X_2, I_2) \in (2^{\mathcal{S}} \setminus \{\emptyset\}) \times (2^{\mathcal{I}} \setminus \{\emptyset\})$, the following condition is satisfied.*

If $X_1 \subseteq X_2$ and $I_1 \sqsubseteq I_2$, then $\phi(X_1, I_1) \geq \phi(X_2, I_2)$.

This condition specifies that a quality function is anti-monotonic with respect to object sets and information sets. Indeed, the definition of 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 $I_1 \sqsubseteq I_2$, the pair (X_1, I_1) is deduced from (X_2, I_2). Therefore, if the (X_2, I_2) is viewed with quality $\phi(X_2, I_2)$, then (X_1, I_1) is viewed with at least the quality $\phi(X_2, I_2)$. In particular, for any (X, I), $\phi(X, I) \leq \phi(X, \{\bot\})$.

A direct consequence of the condition in the Definition 7 is:

If $X_1 = X_2$ and $I_1 \simeq I_2$ then $\phi(X_1, I_1) = \phi(X_2, I_2)$.

Further, if $\mathcal{D}_1$ and $\mathcal{D}_2$ are equivalent information sources, with the source mappings f_1 and f_2, respectively, then for any $x \in \mathcal{S}$, $\phi(\{x\}, f_1(x)) = \phi(\{x\}, f_2(x))$.

Definition 8 *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 Q-*quality view *of $\mathcal{D}$ is a function φ from R into Q, such that for any $(X_1, I_1), (X_2, I_2) \in R$ the following conditions are satisfied:*
(i) If $X_1 \subseteq X_2$ and $I_1 \sqsubseteq I_2$, then $\varphi(X_1, I_1) \geq \varphi(X_2, I_2)$.
(ii) If $I_1 = \{\bot\}$, then $\varphi(X_1, I_1) = max(Q)$.

EXAMPLE 2 We consider the information source $\mathcal{D}_1$ of Example 1. Let $Q = \{A, B, C, D, E\}$ be a quality domain, where $A > B > C > D > E$. For notational convenience, sets are denoted by juxtaposition. For example, $\{1, 4\}$ is denoted by 14, and $\{b, d, e\}$ is denoted by bde. The following tableau defines the Q-quality view φ_1 of $\mathcal{D}_1$.

R_1	(1,adef)	(2,abe)	(3,bde)	(14,af)	(23,be)
φ_1	A	B	B	A	D

Consider $\varphi_1(2, abe)$ and $\varphi_1(23, be)$. Pair $(23, be)$ has an object redundancy with respect to pair $(2, abe)$, because $be \sqsubseteq abe$ and $\{2, 3\} \cap \{2\} \neq \emptyset$. Therefore, in reducing $\mathcal{D}$ into an equivalent information source, pair $(23, be)$ is reduced to pair $(3, be)$. Now, as $\varphi_1(23, be) = D$, we may extend φ_1 by defining $\varphi_1(3, be) = D$. However, if we do so, such an extension of φ_1 is not a Q-quality, because $\varphi_1(3, bde) = B > D = \varphi_1(3, be)$ and $bde \simeq be$.

Definition 9 *Let φ be a Q-quality view of $\mathcal{D}$. A Q-quality extension of φ is a Q-quality, denoted by φ', such that*
(i) For any $(X_i, I_i) \in R$, $\varphi'(X_i, I_i) = \varphi(X_i, I_i)$, and
(ii) For any $(X, I) \in (2^{\mathcal{S}} \setminus \{\emptyset\}) \times (2^{\mathcal{I}} \setminus \{\emptyset\})$, if $I \neq \{\bot\}$ and $\varphi'(X, I) = m \in Q, m > min(Q)$, then there exists $(X_i, I_i) \in R$ such that $X \cap X_i \neq \emptyset$, $I \sqsubseteq I_i$, and $m \leq \varphi(X_i, I_i)$.

Condition (i) requires that φ' must be a correct extension of φ. Condition (ii) requires that for any (X, I) which is not in R, the quality estimation $\varphi'(X, I)$ must be founded on the quality estimation existing for pairs in R.

Now, we show the Q-quality extension of a Q-quality view of $\mathcal{D}$ exists, and how to obtain it.

Proposition 1 *Let $\mathcal{D} = (P(\mathcal{S}), C(\mathcal{I}), R)$ be an information source, and Q be a quality domain. Let φ be a Q-quality view of $\mathcal{D}$. Then φ can be extended to a Q-quality of $\mathcal{D}$.*

For the proof of Proposition 1, we build an extension of φ, denoted by φ' satisfying Definition 7. Consider a pair $(X, I) \in 2^{\mathcal{S}} \times 2^{\mathcal{I}}$, such that $(X, I) \neq (\emptyset, \emptyset)$. 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 $X \cap X_i \neq \emptyset$ and the union of all such $X \cap X_i$ is equal to X. We have the following cases:
Case 1: If $I = \{\perp\}$, then define $\varphi'(X, I) = max(Q)$.

Case 2: Else, if there exists $(X_i, I_i) \in R$ such that $X \subseteq X_i$ and $I \sqsubseteq I_i$, then define $\varphi'(X, I)$
$= max\{\varphi(X_i, I_i) \mid (X_i, I_i) \in R, X \subseteq X_i, I \sqsubseteq I_i\}$

Case 3: Else, let $K = \otimes\{I_i \mid (X_i, I_i) \in R, X \cap X_i \neq \emptyset\}$.

Case 3.1: If $I \sqsubseteq K$, then define $\varphi'(X, I)$
$= min\{m(X_i, I_i) \mid (X_i, I_i) \in R, X \cap X_i \neq \emptyset\}$,
where for each $(X_i, I_i) \in R$ such that $X \cap X_i \neq \emptyset$, $m(X_i, I_i) = max\{\varphi(X_j, I_j) \mid (X_j, I_j) \in R, X \cap X_j = X \cap X_i, I \otimes I_j = I \otimes I_i\}$

Case 3.2: Else, define $\varphi'(X, I) = min(Q)$.

We can show that the so defined φ' is a well-defined function satisfying the condition specified in Definition 7 and the conditions of Definition 9. Thus, φ' is a Q-quality extension of φ.

Moreover, φ' satisfies further interesting properties.

Proposition 2 *For any Q-quality φ'' which extends a Q-quality view φ of $\mathcal{D}$, for any $(X, I) \in (2^{\mathcal{S}} \setminus \{\emptyset\}) \times (2^{\mathcal{I}} \setminus \{\emptyset\})$, we have*
(i) If (X, I) is in Case 1, then $\varphi''(X, I) \leq \varphi'(X, I)$.
(ii) If (X, I) is in Case 2, then $\varphi''(X, I) \geq \varphi'(X, I)$.
(iii) If (X, I) is in Case 3.1, then $\varphi''(X, I) \leq \varphi'(X, I)$.
(iv) If (X, I) is in Case 3.2, then $\varphi''(X, I) \geq \varphi'(X, I)$.

Apart from points (i) and (iv) of Proposition 2, which deals with special cases of information (respectively, no information or exceeding information), points (ii) and (iii) of Proposition 2 state interesting property of the Q-quality extension of φ, defined in Proposition 1: the extension is optimized. Indeed, when the operator max is used for Case 2 in the proof Proposition 1, every possible Q-quality extension of φ cannot be less than φ' in this case, and when the operator min is used for Case 3.1 in the proof of Proposition 1, every possible Q-quality extension of φ cannot be greater than φ' in this case.

EXAMPLE 3 *Consider the quality-view φ_1 of $\mathcal{D}_1$ in Example 2. The Q-quality extension φ'_1 of φ_1, as defined in the proof of Proposition 1, applied on the reduction $\mathcal{D}'_1$ of $\mathcal{D}_1$, is:*

R'_1	*(14,f)*	*(2,ab)*	*(3,b)*
φ'_1	*A*	*B*	*B*

3.2 Quality based approaches to integration

Now, let $\mathcal{D}_1$ and $\mathcal{D}_2$ be two information sources to be integrated. Let $\varphi_i,\ i = 1, 2$, be quality views of $\mathcal{D}_i$, on a same domain quality Q. Let φ'_i be the Q-quality extension of φ_i. The quality based approach to integrating $\mathcal{D}_1$ and $\mathcal{D}_2$, with respect to the quality domain Q can be defined as follows:

Let $\mathcal{D}'_1$ and $\mathcal{D}'_2$ be the reductions of $\mathcal{D}_1$ and $\mathcal{D}_2$, respectively, and $\varphi_i^{rd},\ i = 1, 2$ be the restriction on R'_i of Q-quality extension φ'_i.

Let φ_0 denote the function defined on the pairs $(X_1 \cap X_2, I_1 \theta I_2)$ resulting in step 3 of algorithm θ-integrate:

(i) For the pessimistic integration:
If $I_1 \otimes I_2 = \{\perp\}$, then $\varphi_0(X_1 \cap X_2, I_1 \otimes I_2) = max(Q)$.
Else,
$\varphi_0(X_1 \cap X_2, I_1 \otimes I_2) = max\{\varphi_1^{rd}(X_1, I_1), \varphi_2^{rd}(X_2, I_2)\}$.

(ii) For the optimistic integration:
$\varphi_0(X_1 \cap X_2, I_1 \oplus I_2) = min\{\varphi_1^{rd}(X_1, I_1), \varphi_2^{rd}(X_2, I_2)\}$.

Lemma 1 *The above defined φ_0 is Q-quality view.*

Proof. We use the notations given in algorithm θ-Integrate. For any pairs $(X_1, I_1), (X_2, I_2)$ in $\mathcal{D}'_1$, we have $X_1 \cap X_2 = \emptyset$. Similarly, for any pairs $(Y_1, J_1), (Y_2, J_2)$ in $\mathcal{D}'_2$, we have $Y_1 \cap Y_2 = \emptyset$. Hence, for any pairs $(Z_1, K_1), (Z_2, K_2)$ resulting in step 3 of algorithm θ-integrate, which can be represented by $(Z_1, K_1) = (X_1 \cap Y_1, I_1 \theta J_1)$ and $(Z_2, K_2) = (X_2 \cap Y_2, I_2 \theta J_2)$, we have $Z_1 \cap Z_2 = \emptyset$, because $X_1 \cap Y_1 \cap X_2 \cap Y_2 = \emptyset$. Thus, in any case condition (i) of Definition 8 is satisfied, since $Z_1 \not\subseteq Z_2$ and $Z_2 \not\subseteq Z_1$. Now, if $K_1 \otimes K_2 = \{\perp\}$, then as defined above $\varphi_0(Z_1 \cap Z_2, K_1 \otimes K_2) = max(Q)$. Hence, condition (ii) of of Definition 8 is satisfied. Thus, the lemma is proved. $\diamond$

We define the quality view for the pessimistic integration (or optimistic integration), denoted by φ, to be the restriction of the Q-quality extension (see Proposition 1) of φ_0 on the reduction of $\mathcal{D}$ resulting in step 4 of algorithm θ-integrate.

Proposition 3 *The function φ associated with the pessimistic (or optimistic) integration as defined above is a Q-quality view of the pessimistic (respectively, optimistic) integration.*

Proof. Let φ be the restriction of the Q-quality extension of φ_0, on the reduction of $\mathcal{D}$ resulting in step 4 of algorithm θ-integrate. As the Q-quality extension of φ_0 is a Q quality, the condition on antimonotonicity is satisfied. In consequence, φ satisfies condition (i) of Definition 8. In particular, for any (Z, K) in the reduction of $\mathcal{D}$, if $K = \{\perp\}$, then $\varphi(Z, K) = max(Q)$, after the definition of φ_0. Thus, φ is a Q-quality view of the pessimistic (respectively, optimistic) integration. $\diamond$

EXAMPLE 4 We consider the integration of the sources $\mathcal{D}_1$ and $\mathcal{D}_2$ in Example 1. Let φ_1 be the Q-quality view of $\mathcal{D}_1$ as given in Example 2:

R_1	(1,adef)	(2,abe)	(3,bde)	(14,af)	(23,be)
φ_1	A	B	B	A	D

Let φ_2 be the Q-quality view of $\mathcal{D}_2$ defined as follows:

R_2	(1,bc)	(2,ae)	(3,ab)	(13,b)	(34,de)
φ_2	B	A	B	B	C

The Q-quality views of the reductions $\mathcal{D}'_1$ and $\mathcal{D}'_2$ (the restrictions of Q-quality extensions of R_1 and R_2 on R'_1 and R'_2) are respectively:

R'_1	(14,f)	(2,ab)	(3,b)
φ'_1	A	B	B

R'_2	(1,bc)	(2,ae)	(3,ab)	(4,de)
φ'_2	B	A	B	C

For the pessimistic integration, the function φ_0 is:

R_{pes}	(1,b)	(2,ae)	(3,b)	(4,de)
φ_0	A	A	B	A

Now, the Q-quality extension (see the proof of Proposition 1) of φ_0, restricted on the reduction of $\mathcal{D}$, denoted by φ, for the pessimistic integration is:

R'_{pes}	(13,b)	(2,ae)	(4,de)
φ	B	A	A

For the optimistic integration, the function φ is:

R_{opt}	(1,$\top$)	(23,ab)	(4,f)
φ_0	B	B	C

We observe that in the pessimistic integration, the resulting information is less specific, but with high quality, whereas in the optimistic integration, the resulting information is more specific, but with lower quality.

3.3 The quality based integration

Let $\mathcal{D}'_1$ and $\mathcal{D}'_2$ be the reductions of the information sources $\mathcal{D}_1$ and $\mathcal{D}_2$, respectively, with quality views $\varphi_i,\ i = 1, 2$. Let $\varphi_i^{rd},\ i = 1, 2$ be the Q-quality view of $\mathcal{D}'_i$, which is the restriction on R'_i of the Q-quality extension of φ_i. Let φ_0 be the function defined on the pairs $(X_1 \cap X_2, I_1\theta I_2)$ resulting in step 3 of algorithm θ-integrate. In the *quality based integration* of $\mathcal{D}_1$ and $\mathcal{D}_2$ we denote such a pair by: $(X_1 \cap X_2, I_1\Omega I_2)$, where $I_1\Omega I_2$ is defined as follows:

(i) If $\varphi_1^{rd}(X_1, I_1) = \varphi_2^{rd}(X_2, I_2)$ then $I_1\Omega I_2 = I_1 \otimes I_2$, with $\varphi_0(X_1 \cap X_2, I_1\Omega I_2) = \varphi_1^{rd}(X_1, I_1)$.

(ii) If $\varphi_1^{rd}(X_1, I_1) < \varphi_2^{rd}(X_2, I_2)$ then $I_1\Omega I_2 = I_2$, with $\varphi_0(X_1 \cap X_2, I_1\Omega I_2) = \varphi_2^{rd}(X_2, I_2)$.

(iii) If $\varphi_1^{rd}(X_1, I_1) > \varphi_2^{rd}(X_2, I_2)$ then $I_1\Omega I_2 = I_1$, with $\varphi_0(X_1 \cap X_2, I_1\Omega I_2) = \varphi_1^{rd}(X_1, I_1)$.

Lemma 2 *The above defined function φ_0 is a Q-quality view of the information source resulting in step 3 of algorithm θ-integrate, where θ is the quality based integration.*

In what follows, the quality based integration is denoted by Ω-integration in short. We defihe quality view for Ω-integration, denoted by φ, to be the restriction of φ_0 on the reduction of $\mathcal{D}$ resulting in step 4 of algorithm θ-integrate.

Proposition 4 *The quality based integration is an intermediate integration, and the defined function φ is a Q-quality view of the result.*

EXAMPLE 5 We consider the quality based integration of the same sources $\mathcal{D}_1$ and $\mathcal{D}_2$ as in Example 4, where the quality views of their reductions are respectively:

R'_1	(14,f)	(2,ab)	(3,b)
φ'_1	A	B	B

R'_2	(1,bc)	(2,ae)	(3,ab)	(4,de)
φ'_2	B	A	B	C

For the quality based integration, the function φ_0 is:

R_Ω	(1,f)	(2,ae)	(3,b)	(4,f)
φ_0	A	A	B	A

The Q-quality extension (see the proof of Proposition 1) of φ_0, restricted on the reduction of the quality based integration, denoted by φ, is:

R'_Ω	(14,f)	(2,ae)	(3,b)
φ	A	A	B

The pessimistic integration and quality based integration privilege quality. However, while the pessimistic integration always searches for the consensus of information, the quality based integration does not. Notice that, with Example 5, in the quality based integration point 1 is associated with f and quality A, while in the pessimistic integration, point 1 is associated with b and quality B. This is due to loss of quality when regrouping points in the process of reduction. Indeed, the function φ_0 of the pessimistic integration associated point 1 with b and quality A.

4 A case study: Land cover application

We apply our approach to fusion the land cover assessments. Land cover information consists of basic information describing the land cover classes (see Figure 2) that occupy the territory, and its assessments via the cartographic representations. On the same basic information there may exist different assessments. Our problem is to fusion different assessments to obtain a better one. In this case study we consider the assessments on LCM2000 [1].

The basic information is described by a list of tuples (*Parcel#, PerPixList,...*) where:

Parcel#: A unique identifier code for each parcel showing occurrence surface in a 100km square.

PerPixList: A list of area percentages of the top five spectral subclasses recorded by satellite images within *Parcel#*. For example, a *PerPixList* as Ab_a,75:Aw_c,20:Gi_b,4:Us_c,1 means the parcel is covered by 75% Barley, 20% Wheat, 4% intensive, 1% suburban (see Figure 2).

The existing assessment of LCM2000 is based on *Broad Habitats* classes (BH). The descriptions of BHs (see Figure

[1] http://www.ceh.ac.uk/data/lcm/LCM2000.shtm

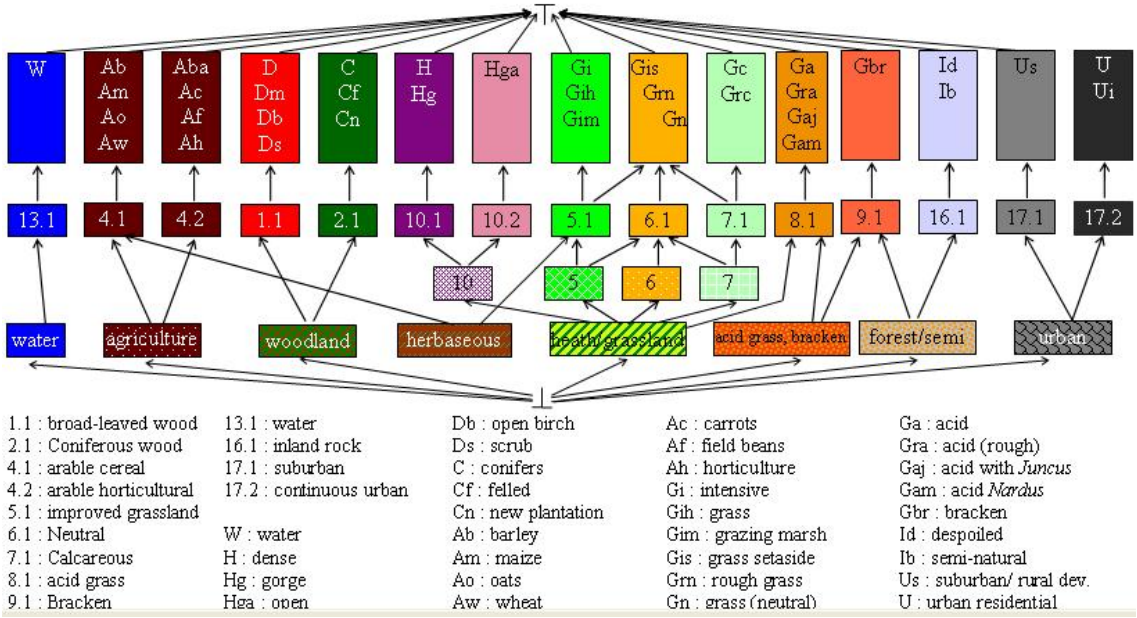


Fig. 2: Information lattice of land cover

2) was developed by the Joint Nature Conservation Committee. LCM2000 aimed to contribute to the assessment of habitats by mapping, as far as possible, the widespread examples of terrestrial, freshwater and coastal Broad Habitats [3]. BHs are complemented by subclasses. In general, we can consider that BHs are defined from dominant subclasses. However, [1] remark that there exists parcels, where the attributed BHs are not the generalization of the dominant subclasses in the parcel. Moreover, the attributed BHs may correspond to non-dominant subclasses.

We propose another method for assessment: the BH of a parcel is the broad habitat class that corresponds to the first subclass in the "PerPixList". Notice that the "PerPixList" is sorted in descending order of the area percentage. This method of assessment can have a drawback: the broad habitat class corresponding to the dominant subclass may not be the dominant broad habitat class. For example, in Table 1, let column BH1 denote the result of the existing assessment of LCM2000, and column BH2 the result of our approach to the assessment of LCM2000. Consider the first line, following our approach, the BH2 value is 2.1 (corresponding to the subclass C_b, with percentage 24). However, the subclasses Dm_a and D_c summing up to 42% correspond to the dominant broad habitat class 1.1 (see Figure 2). Now, consider the second line of Table 1, normally BH1 must be 1.1, because the dominant subclasses are D_a and D_c, with 54 as the sum of percentages (see Figure 2). But the existing assessment of LCM2000 estimated this broad habitat values as 2.1. Thus, both methods are not perfect.

NoP	PerPixList	BH1	BH2
1	C_b,24:Dm_a,22:D_c,20:Ab_k,9	1.1	2.1
2	D_a,36:C_a,34:D_c,18:Am_g,4	2.1	1.1
3	Ac_k,25:Ga_a,17:Gi_e,17	4.2	4.2
...	...	...	...

Table 1: Two methods for assessment

4.1 A data quality measure

In order to evaluate the quality of these methods of assessment, we consider:

The Completeness of basic information: The sum of all percentages of non-null values in the PerPixlist. The smaller sum of percentages in PerPixList corresponds to the lower quality of the estimated broad habitat value.

The Consistency: For each PerPixList, we regroup the subclasses following the hierarchical structure. For example, in the first line of Table 1, subclasses Dm_a and D_c are regrouped into class 1.1 with percentage 42, C_b is regrouped into class 2.1 with percentage 24, and Ab_k is regrouped into class 4.1 with percentage 9 (see Figure 2). The resulting list is sorted in descending order of percentages, called *TopList*: 1.1, 42: 2.1, 24: 4.1,9. We say that an estimated broad habitat value is consistent if it is the first value of the TopList. Hence, for the first line of Tab 1, BH1 is consistent, but not BH2. In other words, BH1 is more consistent then BH2. The worst case is the case where the estimated broad habitat does not correspond to any class in the TopList.

Basing on the above ideas, Table 2 defines a method for assigning quality values of the domain $\{A, B, C, D, E, F\}$ where $A > B > C > D > E > F$.

$Complet.\backslash Consist.$	1^{st}	2^{nd}	3^{rd}	4^{th}	5^{th}
> 90	A	B	C	D	E
> 60	B	C	D	E	F
> 40	C	D	E	F	F
< 40	F	F	F	F	F

Table 2: Method for quality assignment

NoP	Toplist	BH1	BH2	$Cons_1$	$Cons_2$	$Comp_1$	$Comp_2$	φ_1	φ_2	Ω-Integ	φ
1	1.1,42:2.1,24:4.1,9	1.1	2.1	1^{st}	2^{nd}	75	75	B	C	1.1	B
2	1.1,54:2.1,34:4.1,4	2.1	1.1	2^{nd}	1^{st}	92	92	B	A	1.1	A
3	4.2,25:8.1,17:5.1,17	4.2	4.2	1^{st}	1^{st}	59	59	C	C	4.2	C
...	...	...	...	...	...	...	...	...	...	...	...

Table 3: Quality based integration of BH1 and BH2

4.2 Results

The quality based integration of the assements given in Table 1 is resumed in Table 3. We have experimented our approach on the data of LCM2000 [3], for about 2700 parcels. In this section we illustrate the result with a restricted number of parcels in LCM2000. Figures 3, 4, and 5 respectively represent the existing assessment, the PerPixList assessment (Toplist), and the result of the quality based integration, of the same parcels, using the same color code given in Figure 2. On the result of the integration, we observe: The regions where the assessments are in conflict correspond to striped colors. The classes with these striped colors are in fact the consensus of the conflicting classes. In particular, the white color corresponds to total conflict.

The quality of each assessment is given in Figures 6, 7, and 8, respectively. The color code for quality values is as follows: A: green; B: yellow; C: orange; D: red; E: dark red; and F: black.

We can observe that the result of the quality based integration is better than the two assessment sources.

5 Conclusion

We have presented an approach to integration of information sources, where information space has the lattice structure, assuming quality data of information is available in the sources. The main problems we have to deal with concern the redundant, complementary, and/or conflicting data. Different methods for combining information of the sources, using available quality data, are proposed. In particular, the quality based integration method privileges information with best quality. In case of the same quality, consensus solution is proposed for conflicting information. Experimental result shows that the approach is useful and interesting: for each parcel, the BH class of the assessment with the best quality is selected as the result of the integration; When the two assessments assign different classes to the parcels, but with the same quality, then the consensus class is proposed for the result of integration. Conflicting classes are detected on parcels, and consensus classes are shown on the map by striped colors. The quality of the integration result is actually improved, as we can see through the quality maps: the areas with colors black, dark red, red, and orange, respectively, of the last quality map are smaller than the same areas in the first two quality maps. We remark that when increasing quality in the result of integration, the corresponding information generally becomes less specific. However, with this approach it is possible that when increasing quality in the result of integration, the corresponding information may also more specific, as the assessment with better quality may provides the information which is more specific.

Acknowledgements

This work is supported by the European Community, under contract IST-1999-14189: REVIGIS project of Future and Emerging Technologies program of IST, FP5. Data are provided by courtesy of C.E.H [2], through an agreement with Department of Geography, University of Leicester.

References

[1] A.J. Comber, P.F. Fisher and R.A. Wadsworth. Creating Spatial Information: Commissioning the UK Land Cover Map 2000. In Advances in Spatial Data, D. Richardson, and P. van Oosterom, editors. *Berlin: Springer-Verlag*, 351-362. 2002.

[2] F. Fonseca, M. Egenhofeer, C. Davis and G. Câmara. Semantic Granularity in Ontology-Driven Geographic Information Systems. In *AMAI Annals of Mathematics and Artificial Intelligence - Special Issue on Spatial and Temporal Granularity*, 36:121-151, 2002.

[3] R.M. Fuller, G.M. Smith, J.M. Sanderson, R.A. Hill, A.G. Thomson, R. Cox, N.J. Brown, R.T Clarke, P. Rothery and F.F. Gerard. Land Cover Map 2000, a guide to the classification system. Technical report of Centre for Ecology and Hydrology, 2002.

[4] V. Phan Luong, T.T. Pham and R. Jeansoulin. Integrating Information under Lattice Structure. In *Proc. of 14th Int. Symposium on methodologies for intelligent system*, ISMIS03, 83–87, Maebashi City, Japan, October, 2003.

[5] G. Stumme and A. Maedche. FCA-Merge: Bottom-Up Merging of Ontologies. In *Proc. 17th Int. Joint Conf. on Artifical Intelligence*, IJCAI01, 225–234, 2001.

[6] H. Veregin. Data quality parameters. In *the journal of Geographical Information Systems*, vol. 1, 177–189, 1999.

[7] L. Wald. Definitions and terms of reference in data fusion. In *Int. Archives of Photogrammetry and Remote Sensing*, Vol. 32, Part 7-4-3 W6, Valladolid, Spain, June, 1999.

[8] M. Worboys and M. Duckham. Integrating spatio-thematic information. In *Proc. of conference GIScience*, Colorado, USA, 2002.

[9] M. Worboys and E. Clementini. Integration of Imperfect Spatial Information. In *the journal of Visual Languages and Computing*, 12: 61–80, 2001.

[2] http://www.ceh.ac.uk/

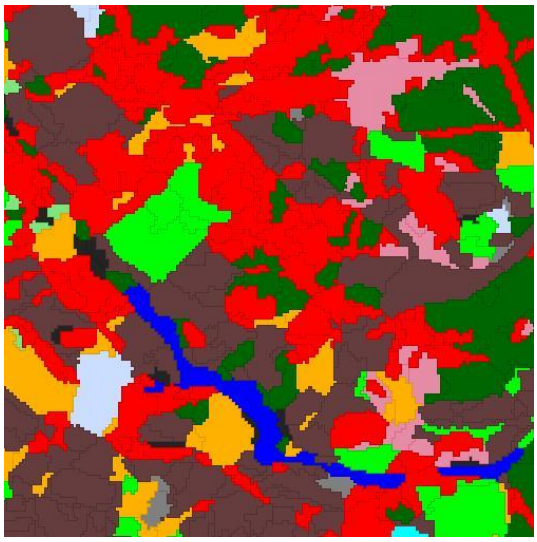


Fig. 3: Existing assessment of LCM2000

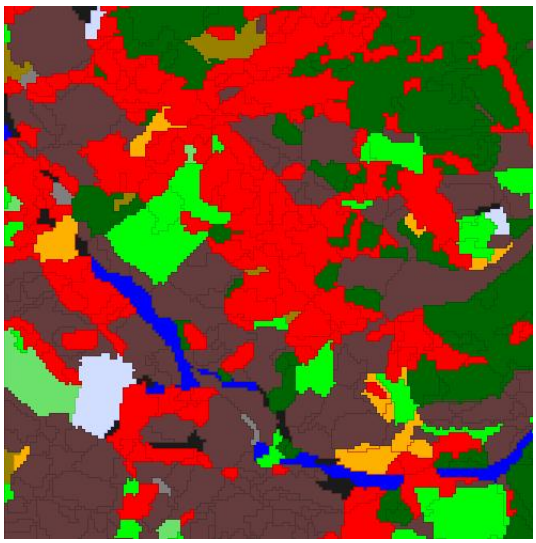


Fig. 4: A new assessment

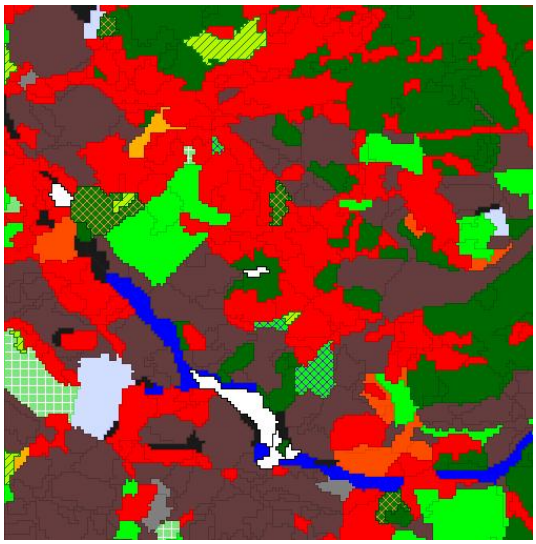


Fig. 5: Result of integration

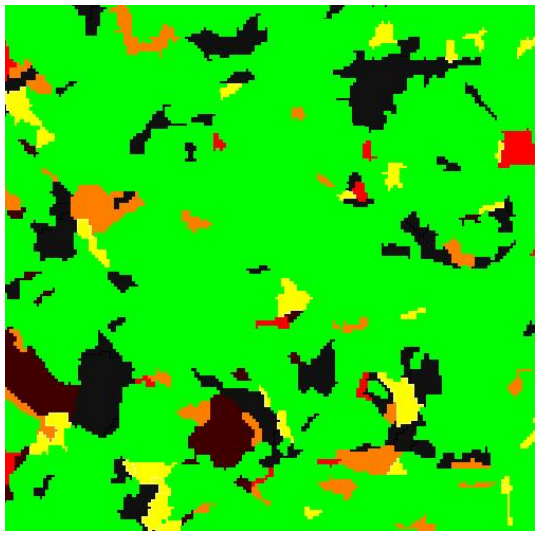


Fig. 6: Quality of the existing assessment

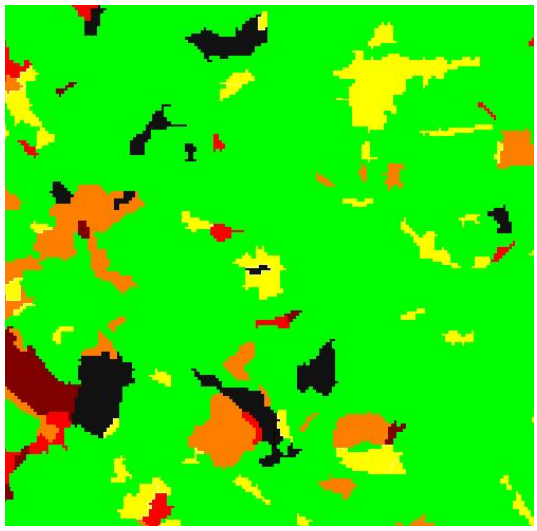


Fig. 7: Quality of the new assessment

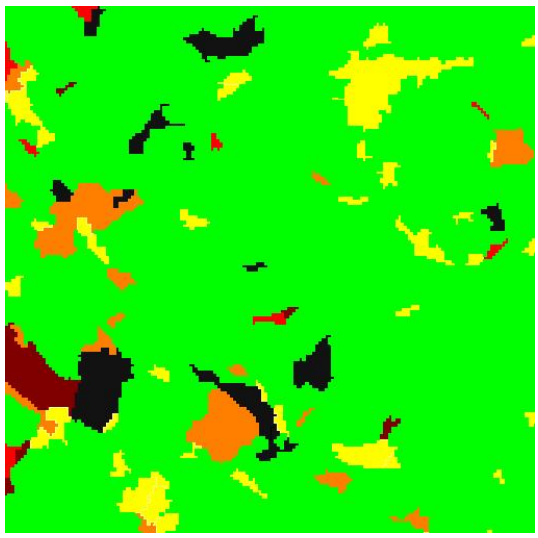


Fig. 8: Quality of integration assessment