Probability Sampling Protocol for Thematic and Spatial Quality Assessment of Classification Maps Generated From Spaceborne/Airborne Very High Resolution Images

Andrea Baraldi, Luigi Boschetti, and Michael L. Humber

***Abstract*—To deliver sample estimates provided with the necessary probability foundation to permit generalization from the sample data subset to the whole target population being sampled, probability sampling strategies are required to satisfy three necessary not sufficient conditions: 1) All inclusion probabilities be greater than zero in the target population to be sampled. If some sampling units have an inclusion probability of zero, then a map accuracy assessment does not represent the entire target region depicted in the map to be assessed. 2) The inclusion probabilities must be: a) knowable for nonsampled units and b) known for those units selected in the sample: since the inclusion probability determines the weight attached to each sampling unit in the accuracy estimation formulas, if the inclusion probabilities are unknown, so are the estimation weights. This original work presents a novel (to the best of these authors' knowledge, the first) probability sampling protocol for quality assessment and comparison of thematic maps generated from spaceborne/airborne very high resolution images, where: 1) an original Categorical Variable Pair Similarity Index (proposed in two different formulations) is estimated as a fuzzy degree of match between a reference and a test semantic vocabulary, which may not coincide, and 2) both symbolic pixel-based thematic quality indicators (TQIs) and sub-symbolic object-based spatial quality indicators (SQIs) are estimated with a degree of uncertainty in measurement in compliance with the well-known Quality Assurance Framework for Earth Observation (QA4EO) guidelines. Like a decision-tree, any protocol (guidelines for best practice) comprises a set of rules, equivalent to structural knowledge, and an order of presentation of the rule set, known as procedural knowledge. The combination of these two levels of knowledge makes an original protocol worth more than the sum of its parts. The several degrees of novelty of the proposed probability sampling protocol are highlighted in this paper, at the levels of understanding of both structural and procedural knowledge, in comparison with related multi-disciplinary works selected from the existing literature. In the experimental session, the proposed protocol is tested for accuracy validation of preliminary classification maps automatically generated by the Satellite Image Automatic Mapper (SIAM™) software product from two WorldView-2 images and one QuickBird-2 image provided by DigitalGlobe for testing purposes. In these experiments, collected TQIs and SQIs are statistically valid, statistically significant, consistent across maps, and in agreement with theoretical expectations, visual (qualitative) evidence and quantitative quality indexes of operativeness (OQIs) claimed for SIAM™ by related papers. As a subsidiary conclusion, the statistically consistent and statistically significant accuracy validation of the SIAM™ pre-classification maps proposed in this contribution, together with OQIs claimed for SIAM™ by related works, make the operational (automatic, accurate, near real-time, robust, scalable) SIAM™ software product eligible for opening up new inter-disciplinary research and market opportunities in accordance with the visionary goal of the Global Earth Observation System of Systems initiative and the QA4EO international guidelines.**

***Index Terms*—Contingency matrix, error matrix, land cover change (LCC) detection, land cover classification, maps comparison, nonprobability sampling, ontology, overlapping area matrix (OAMTRX), probability sampling, quality indicator of operativeness (OQI), spatial quality indicator (SQI), taxonomy, thematic quality indicator (TQI).**

ACRONYMS

ATCOR	Atmospheric/Topographic Correction
B	(visible) Blue
BHR	Bi-Hemispherical Reflectance
BRDF	Bidirectional Reflectance Distribution Function
Cal/Val	Calibration/Validation
CEOS	Committee on Earth Observation Satellites
CVPSI	Categorical Variable Pair Similarity Index
DN	Digital Number
DSM	Digital Surface Model
DTM	Digital Terrain Model
EO	Earth Observation
FIEOS	fourth-generation Future Intelligent Earth Observation Satellites
G	(visible) Green
GEO	Group on Earth Observations
GEOBIA	Geographic Object-Based Image Analysis
GEOOIA	Geographic Object-Observation Image Analysis
GEOROI	Geographic Region Of Interest

Manuscript received June 3, 2012; revised November 13, 2012 and January 3, 2013; accepted January 15, 2013. Date of publication March 13, 2013; date of current version November 26, 2013. This work was supported in part by the National Aeronautics and Space Administration under Grant/Contract/Agreement NNX07AV19G issued through the Earth Science Division of the Science Mission Directorate. The research leading to these results has also received funding from the European Union Seventh Framework Programme FP7/2007-2013 under Grant Agreement 263435 with the project title: BIOdiversity Multi-Source Monitoring System—from Space TO Species (BIO-SOS).

A. Baraldi and M. L. Humber are with the Department of Geographical Sciences, University of Maryland, College Park, MD 20742 USA (e-mail: andrea.baraldi@hermes.geog.umd.edu; mhumber@umd.edu).

L. Boschetti was with the University of Maryland, College Park, MD 20742 USA. He is now with the College of Natural Resources, University of Idaho, Moscow, ID 83844-4264 USA (e-mail: luigi@uidaho.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TGRS.2013.2243739

GEOSS	Global EO System of Systems
GIS	Geographic Information System
GIScience	Geographic Information Science
GMES	Global Monitoring for the Environment and Security
GRNS	Greenness
HR	High spatial Resolution
IR	Infrared
IUS	Image Understanding System
LAI	Leaf Area Index
LC	Land Cover
LCC	Land Cover Change
LCLUC	Land Cover and Land Use Change program
LR	Low spatial Resolution
MIR	Medium-IR
MR	Medium spatial Resolution
MS	Multi-Spectral
NASA	National Aeronautics and Space Administration
OA	Overall Accuracy
OQI	Quality Indicator of Operativeness
QA	Quality Assurance
QA4EO	Quality Accuracy Framework for Earth Observation
QB-2	QuickBird-2
QI	Quality Indicator
R	(Visible) Red
RE	Red Edge
ROI	Region Of Interest
RS	Remote Sensing
RS-IUS	Remote Sensing Image Understanding System
SIAM™	Satellite Image Automatic Mapper
SIRS	Simple Random Sampling
STRS	Stratified Random Sampling
SURF	Surface Reflectance
SWIR	Short Wave Infrared
TIR	Thermal IR
TM	Trademark
TOA	Top-Of-Atmosphere
TOARD	TOA Radiance
TOARF	TOA Reflectance
TOC	Topographic Correction
USGS	US Geological Survey
VHR	Very High spatial Resolution
VIS	Visible (spectral band)
WELD	Web-Enabled Landsat Data set project
WGCV	Working Group on Calibration and Validation
WV-2	WorldView-2

I. Introduction

In RECENT years, the demand for high spatial resolution (HR, ranging from 20 to 5 m) and very HR (VHR, below 5 m) spaceborne/airborne imagery has continued to increase in terms of data quantity and quality, which has boosted the rapid growth of the commercial VHR satellite industry [1]. The ever-increasing accessibility/availability of spaceborne/airborne VHR images represents a major challenge for those portions of the remote sensing (RS) community involved with the development of satellite-based information/knowledge processing systems [2]. In the Quality Assurance Framework for Earth Observation (QA4EO) guidelines, delivered by the international Group on Earth Observations (GEO) Committee on Earth Observation Satellites (CEOS) and adopted by the Global Earth Observation System of Systems (GEOSS) implementation plan for years 2005–2015 [3], [4], satellite-based information/knowledge processing systems are required to be suitable/reliable to allow the provision of "*the Right Information, in the Right Format, at the Right Time, to the Right People, to Make the Right Decisions*."

Unfortunately, to date, the automatic or semi-automatic transformation of huge amounts of multi-source multi-resolution Earth observation (EO) images into information/knowledge can still be considered far more problematic than might be reasonably expected [5]–[19]. In practice, the increasing rate of collection of EO images of enhanced spatial, spectral, and temporal quality outpaces the ability of existing RS image understanding systems (RS-IUSs, where terms "*image understanding system*" and "*computer vision system*" are synonyms) to infer from sensory data: 1) either continuous or discrete sub-symbolic (e.g., biophysical) variables, e.g., a leaf area index (LAI) map, or 2) (discrete) categorical variables, e.g., land cover (LC) and LC change (LCC) maps.

Possible causes of the ongoing lack of operational (turnkey, ready-to-go, good-to-go) RS-IUS solutions can be investigated at the *four levels of understanding of an information processing system*, namely, 1) computational theory (system architecture), 2) knowledge/information representation, 3) algorithms, and 4) implementation. Existing literature clearly acknowledges that the linchpin of success of an information processing system is addressing the system design and information/knowledge representation, rather than algorithms or implementation [5]–[17], [20]–[22]. This implies that, since existing RS-IUSs encompass a huge variety of algorithms and implementations, then their ongoing lack of productivity should be investigated at the levels of understanding of computational theory and knowledge/information representation [5]–[17]. If this conjecture holds true then, to achieve the visionary goal of international initiatives like the ongoing GEO GEOSS project [3], [4], a new generation of operational RS-IUSs, based on enhanced multi-disciplinary foundations in computational theory and information representation, should be developed by the RS community in close cooperation with academic disciplines like philosophical hermeneutics [23], [24], geography [25], [26], neurophysiology [27]–[30], psychophysics [31], machine learning [32]–[34], artificial intelligence [35]–[37], computer vision [21], and computer science [26].

When dealing with spaceborne/airborne VHR imagery, the inadequacy of the RS community to provide operational RS-IUS solutions becomes twofold. On the one hand, the spatial and spectral resolution of spaceborne/airborne VHR imagery makes *the complexity of VHR image understanding analogous to the degree of complexity of human vision, recognized as a hybrid (combined deductive and inductive) cognitive problem* [5]–[17], [35], [38]–[45], *inherently ill-posed in the Hadamard sense* [46] *and, therefore, very difficult to solve* [32]. In the words of Iqbal and Aggarwal: "Frequently, no claim is made

about the pertinence or adequacy of the digital models as embodied by computer algorithms to the proper model of human visual perception... This enigmatic situation arises because research and development in computer vision is often considered quite separate from research into the functioning of human vision. A fact that is generally ignored is that biological vision is currently the only measure of the incompleteness of the current stage of computer vision and illustrates that the problem is still open to solution" [47].

In addition to the complexity of VHR image understanding, there is an incapacity of the RS community to validate VHR image-derived products for the assessment and comparison of competing RS-IUS solutions. This means that *the validation of VHR image-derived products is, per se, a cognitive problem whose complexity, related to the development of RS-IUSs capable of VHR image understanding, should be considered of the same order of complexity of human vision*.

The primary objective of this work is to propose to the RS community a novel (to the best of these authors' knowledge, the first) probability sampling protocol for the accuracy assessment (*accuracy validation* [2], [3], [48]) and comparison of classification maps generated from spaceborne/airborne VHR images, where complementary symbolic pixel-based thematic quality indicators (TQIs) [49]–[51] and sub-symbolic polygon-based spatial quality indicators (SQIs) [52]–[54] are estimated with a degree of tolerance in measurement in compliance with the QA4EO international guidelines [3]. It is worth mentioning that, like a decision-tree, any *protocol* (guidelines for best practice [48]) comprises a set of rules, equivalent to *structural knowledge*, and an order of presentation of the rule set, called *procedural knowledge*. The combination of these two levels of knowledge makes an original protocol worth more than the sum of its (eventually non-original) parts.

As a subsidiary objective of this work, the proposed protocol is tested for accuracy validation of preliminary classification maps (pre-attentive vision classification maps, pre-classification maps) generated, by means of the Satellite Image Automatic Mapper (SIAM™) software toolbox selected from the existing literature [5]–[17], from a set of multi-source, multi-resolution, multi-temporal VHR images provided by DigitalGlobe for testing purposes. Presented in recent years to the RS community for operational use in a RS-IUS pre-attentive vision first stage, SIAM™ accomplishes multi-scale image segmentation and multi-granularity image pre-classification simultaneously, automatically and in near real-time [5]–[17].

The several degrees of novelty of the proposed probability sampling protocol encompass both levels of structural and procedural knowledge as summarized below.

I) To the best of these authors' knowledge, this is the first time the probability sampling design of Stehman and Czaplewski [55] is instantiated to assess and compare thematic maps generated from VHR images, where two sets of uncorrelated symbolic pixel-based TQIs [49]–[51] and sub-symbolic object-based SQIs [52]–[54], [56] are taken into account under the following constraints.
   - Every metrological/statistically-based Quality Indicator (QI) must be provided with a variance estimation to be considered statistically significant, in compliance with principles of statistics and the QA4EO guidelines [3], [55].
   - QI ensembles (e.g., TQIs, SQIs) have to account for the well-known non-injective property of QIs [5]–[17]. It is common knowledge that, given any QI formulation, it is always possible to find two different target phenomena featuring the same QI value, e.g., two different thematic maps may feature the same overall accuracy (OA) index [50], [51]. This implies that no hypothetical universal QI can exist, which contradicts a significant portion of the RS literature [57]–[59]. This is tantamount to saying that, in any quality assessment task, a set of mutually uncorrelated QIs has to be carefully selected by domain experts to account for the non-injectivity of QIs [5]–[17].

It is worth mentioning that, according to Stehman and Czaplewski [55], probability or nonprobability sampling protocols for map accuracy assessment consist of six basic components.

1) Identification of the test map taxonomy, reference sample set taxonomy, and their *contingency table* (*error matrix*). In general, semantic associations between legend pairs are many-to-many, whose special cases are many-to-one, one-to-many, and one-to-one relations.
2) The sampling design protocol, by which sampling units are selected into the sample.
3) The evaluation protocol, to collect information contributing to the reference classification determination.
4) The labeling protocol, to assign the reference classification(s) to the sampling units based on the information collected by the evaluation protocol.
5) The analysis protocol, where a contingency table is instantiated.
6) The estimation protocol, where QIs (summary statistics, summary measures [55]) are collected from the contingency table(s) and assessed in comparison with reference standards.

By definition, probability sampling must satisfy three necessary not sufficient conditions to deliver statistically valid (consistent) sample estimates, i.e., sample estimates provided with the necessary probability foundation to permit generalization from the sample data subset to the whole target population being sampled [55], [60].

(i) All inclusion probabilities be greater than zero in the target population to be sampled. If some sampling units have an inclusion probability of zero, then the accuracy assessment does not represent the entire target region depicted in the map to be assessed.
(ii) The inclusion probabilities must be [60]:
   - knowable for nonsampled units and
   - known for those units selected in the sample: since the inclusion probability determines the weight attached to each sampling unit in the accuracy estimation formulas, if the inclusion probabilities are unknown, so are the estimation weights.

TABLE I
EXISTING COMMERCIAL RS-IUS SOFTWARE PRODUCTS AND THEIR DEGREE OF MATCH WITH THE INTERNATIONAL QA4EO GUIDELINES [5]

RS-IUS Commercial software products	Output of a pre-attentive vision first stage (e.g. image segmentation, image pre-classification, etc.), if any, in a RS-IUS software product: Sub-symbolic (asemantic) versus Symbolic (semantic) information primitives, whose spatial type is either Pixel, Polygon (Segment) or Multi-Part Polygon (Stratum, Layer)	Radiometric Calibration (RAD. CAL.) requirement in compliance with the international QA4EO guidelines [3], where RAD. CAL.is considered mandatory from sensor building to end-of-life to ensure RS data harmonization and interoperability
PCI Geomatics GeomaticaX	Sub-symbolic pixels	RAD. CAL. is not required, but optional. It means this a statistical model-based RS-IUS, inherently semi-automatic and site-specific.
Trimble eCognition Developer [89], [90]	Unsupervised data learning sub-symbolic polygons	RAD. CAL. is not required, but optional. It means this a statistical model-based RS-IUS, inherently semi-automatic and site-specific.
Pixel- and Segment-based versions of the Environment for Visualizing Images (ENVI) by ITT VIS [141]	Either sub-symbolic pixels or unsupervised data learning sub-symbolic polygons	RAD. CAL. is not required, but optional. It means this a statistical model-based RS-IUS, inherently semi-automatic and site-specific.
ERDAS IMAGING Objective	Supervised data learning symbolic polygons	RAD. CAL. is not required, but optional. It means this a statistical model-based RS-IUS, inherently semi-automatic and site-specific.
Atmospheric/Topographic Correction-2/3/4 (ATCOR-2/3/4) [83]-[86]	Sub-symbolic pixels or symbolic pixels, where the semantic label, if any, is a spectral type provided by the physical model-based Spectral Classification of surface reflectance signatures (ATCOR-SPECL) by-product.	Physical model-based RS-IUS, consistent with the QA4EO recommendations: RAD. CAL. into surface reflectance (SURF) values is required $\Rightarrow$ inherently ill-posed atmospheric correction first stage $\Rightarrow$ semi-automatic and site-specific.
Novel three-stage stratified hierarchical hybrid RS-IUS employing the Satellite Image Automatic Mapper (SIAM^{TM}) as its preliminary classification first stage [5]-[17]	Physical model-based symbolic pixels $\in$ symbolic polygons $\in$ symbolic multi-part polygons	Physical model-based RS-IUS, consistent with the QA4EO recommendations: RAD. CAL. into top-of-atmosphere (TOA) reflectance (TOARF) or surface reflectance (SURF) values, with TOARF $\approx$ (SURF + atmospheric noise) $\supset$ SURF, is required $\Rightarrow$ atmospheric correction is optional. Automatic and robust to changes in RS optical imagery acquired across time, space and sensors.

II) The aforementioned project requirements specification means that:
- the proposed probability sampling protocol is *statistically valid*, i.e., *consistent*, so that the sampling represents the entire target region of a test map [55].
- Since they are provided with a confidence interval, estimated TQIs and SQIs are *statistically significant*.

These project requirements are neither trivial nor obvious. In the existing RS literature, e.g., [54], [61], [62], both segmentation and classification map accuracies are typically estimated via nonprobability sampling, where inclusion probabilities of selected samples are unknown or ignored, while QIs are not provided with any degree of uncertainty in measurement, in violation with the principles of statistics and the QA4EO guidelines [3].

III) Stehman describes four common types of maps comparison [63]. In the first type, different thematic maps, either crisp of fuzzy [64], [65], of the same region and employing the same sorted set (legend) of LC classes, are compared [66]. To date, a large segment of the RS community appears concerned with this first type of maps comparison exclusively [51], [66]. In the second type, which includes the first type as a special case, thematic maps, either crisp of fuzzy, of the same region, but featuring map legends that differ in semantics and/or cardinality (size) and/or order of presentation, are compared [63]. The present work focuses on this second type of inter-map comparisons.

IV) Following identification of the test map and reference sample set taxonomies (refer to point I.1 above in this section), a *contingency table* (*error matrix*) must be selected. In this paper, where test and reference semantic vocabularies may not be the same (refer to point III above in this section), the error matrix becomes an either square or non-square *overlapping area matrix* (OAMTRX) [52], [67]. The concept of OAMTRX, found in literature [63], [68], is a generalization of the well-known concept of (square and sorted) *confusion matrix* (CMTRX) [50], [51].

V) Following identification of the test map and reference sample set taxonomies (refer to point I.1 above in this section), an OAMTRX instance, either square or non-square, must be defined by a knowledge expert (knowledge engineer [69]) who selects table entries (test-reference class pairs) to be considered as "*correct*". These "*correct*" entries may be diagonal or off-diagonal OAMTRX cells. The matching between two legends is, *per se*, a cognitive (interpretation) process whose "*information-as(an interpretation)-process*" is inherently equivocal [23], [24]. Hence, categorical variable matching should require negotiation and be community agreed [68], [70], [71]–[76]. Unlike the interpretation of a CMTRX, where the main diagonal guides the interpretation process (at least in terms of overal accuracy estimation [49]–[51]), comprehensive interpretation of an OAMTRX instance can be very challenging, complex and time consuming because,

in general, a non-square OAMTRX has no main diagonal to guide the interpretation process or, if the OAMTRX is square, its main diagonal may not consist, in part or at all, of "*correct*" entries [52], [67], [68].

VI) Following definition of an OAMTRX (refer to point V above in this section), an original QI, identified as the Categorical Variable Pair Similarity Index ($CVPSI$) $\in$ [0, 1], is estimated as a *fuzzy degree of match between the reference and the test semantic vocabulary*. In practice, $CVPSI$ is a fuzzy degree of similarity between an OAMTRX definition, where "*correct*" (allowed) reference-test class relations are, in general, many-to-many, with an (ideal) CMTRX, where allowed reference-test class relations are one-to-one exclusively. *Vice versa*, $(1 - CVPSI)$ is a *normalized estimate of the additional (classification) work required to fill up the semantic gap from the test semantic vocabulary to the reference (target) semantic vocabulary*. For example, in the ideal case of an OAMTRX where only one-to-one reference-test class relations can be found, irrespective of the fact they are diagonal or off-diagonal entries, then condition $CVPSI$ equal to 1 holds.

VII) In the probability sampling phase (refer to point I.2 above in this section), a general rule of thumb would require to select the reference data source "one step closer to the ground" than the RS data used to make up the test map [51]. Unfortunately, when dealing with thematic maps generated from VHR imagery, it is often the case there is no reference data source acquired at the same time of the VHR image and one step closer to the ground. Hence, the sole data source available for reference population sampling is the same VHR image adopted to generate the test map. In other words, *the test and reference data sources coincide with the VHR image at hand*.

VIII) In the probability sampling phase (refer to point I.2 above in this section), no prior knowledge of the class-specific reference strata (layers) is available to run a stratified random sampling (STRS) strategy in the VHR image at hand (refer to point VII above). Hence, an original non-standard simple random sampling (SIRS) procedure is applied per reference class (refer to Fig. 12 in Section VI-B2 below).

IX) In the estimation phase (refer to point I.6 above in this section), symbolic pixel-based TQIs are selected from the existing literature [49], together with their variance estimation formulas [50]. In general, a different variance estimator formula arises for each accuracy index and each different sampling design [55]. In this paper, TQIs, provided with a degree of uncertainty in measurement, are estimated from a non-traditional OAMTRX instance, which was defined and instantiated in the sample analysis phase (refer to point V above in this section).

X) In the estimation phase (refer to point I.6 above in this section), sub-symbolic polygon-based SQIs are selected, augmented and instantiated in line with the RS literature [54], together with their variance estimation formulas [50]. In this substep of the estimation phase, thematic information is ignored in the comparison between test and reference samples, since thematic matching has already been accounted for by symbolic pixel-based TQIs. In practice, a (symbolic) classification map is transformed into a (sub-symbolic) segmentation map.[1] Thus, pairs of test and reference polygons (2-D objects, segments) are compared in terms of shape, irrespective of their thematic labels. This step in sample analysis corresponds to the investigation of the spatial distribution of thematic errors, which has been highly recommended in literature [52], [53], [56], but almost never performed in RS common practice.

XI) In compliance with the definition of probability sampling (refer to this section above), (unequal) inclusion probabilities of sampling units, either pixels or polygons employed in the estimation of, respectively, TQIs and SQIs, are assessed to determine the weight, equal to the inverse of the inclusion probability, attached to each sampling unit in the Horvitz–Thompson sample estimator. The Horvitz–Thompson theorem guarantees that the Horvitz–Thompson sample estimator is unbiased for the population total [60] (for further details about the Horvitz–Thompson theorem, refer to Section V below).

The main experimental conclusion of this work is that the proposed protocol is tested successfully in the accuracy validation of the SIAM™ multi-granularity maps automatically generated from multi-sensor multi-temporal VHR images. In these experiments, collected TQIs and SQIs are statistically valid and statistically significant, consistent across maps, and in agreement with theoretical considerations, visual (qualitative) evidence, and (quantitative) QIs of operativeness (OQIs) claimed for SIAM™ by the existing literature [5]–[17]. Estimated SQIs are found to be negatively biased (e.g., underestimated) due to, first, an eight-adjacency neighborhood effect and, second, an inadequacy to cope with a test and a reference semantic vocabulary when these vocabularies do not coincide.

As a subsidiary conclusion, the statistically consistent and statistically significant accuracy validation of the SIAM™ pre-classification maps proposed in this work, in combination with the high-value OQIs claimed for SIAM™ by related papers [5]–[17], make the operational (fast, accurate, automatic, robust, scalable) SIAM™ software product eligible for opening up new inter-disciplinary research and market opportunities in compliance with the visionary goal of the GEO GEOSS initiative and the GEO QA4EO guidelines.

The rest of this paper is organized as follows. Section II presents definitions, terminology, and multi-disciplinary concepts related to this work. In Section III, problem recognition and opportunity identification are accomplished in a multi-disciplinary framework. Section IV presents the test VHR image set, provided by DigitalGlobe, and discusses the VHR image pre-processing activities carried out before running

[1]The generation of a segmentation map from a binary mask or multi-level image (e.g., a thematic map) is a well-posed segmentation problem (i.e., the problem solution exists and is unique), typically solved by a computationally efficient two-pass connected-component image labeling algorithm [77]. In practice, a unique (sub-symbolic) segmentation map can be generated from a (symbolic) thematic map, but the contrary does not hold, i.e., different thematic maps can generate the same segmentation map [52].

the SIAM™ preliminary classifier. The Horvitz–Thompson theorem is discussed in Section V, where original inclusion probabilities suitable for non-standard sampling designs are proposed. In Section VI, a novel protocol to operationalize the thematic and spatial accuracy assessment of thematic maps generated from spaceborne/airborne VHR images is proposed. In addition, this novel protocol is instantiated to validate the preliminary classification maps automatically generated by the SIAM™ from the VHR test images. In Section VII, OQIs of the SIAM™ software product inferred from Section VI are summarized. Section VIII presents new inter-disciplinary research and market opportunities opened up by the operational, automatic, near real-time SIAM™ pre-attentive vision classifier. Conclusions are reported in Section IX.

II. Adopted Definitions, Terminology, and Multi-Disciplinary Concepts

This section is provided with a significant survey value to make this paper self-contained for a potential readership unfamiliar with the multi-disciplinary background of RS image understanding. In this paper, the following definitions, terminology, and multi-disciplinary concepts hold.

A. Quantitative and Qualitative Concepts of Information

- The following concepts are defined in compliance with philosophical hermeneutics [23], [24].
  - Numerical, sensory, quantitative "*data*" as synonyms of observables, true facts. It is important to stress that sensory data are provided, *per se*, with no semantics at all [19].
  - Sub-symbolic, quantitative, unequivocal "*information-as-thing*," according to the Shannon theory of communication [78]. Quantitative information according to Shannon is an object or a thing (e.g., number of bits, number of words in a document, etc.) irrespective of its meaning. This makes the information exchange between a sender and a receiver unequivocal, therefore easier to deal with than when meaning is involved in the communication process [16], [23], [24].
  - Symbolic, qualitative, equivocal "*information-as-(an intepretation) process*," i.e., *information as interpreted data* [23], [24]. In the words of philosophical hermeneutics, symbolic information is always related to "a receiver's beliefs, desires, and background knowledge," i.e., the meaning of a message is always context dependent. It is indeed a "harmless fiction" to think about a meaning of a message at the source "independently of what anyone (a receiver) happened to know" [23], [24]. There are no inquirers (users, knowers, receivers, cognitive agents) in general, but context-dependent users. Analogously, (objectivized, externalized) information systems (e.g., database systems) are always embedded in various social, cultural, etc., contexts [23], [24].
  - "*Knowledge*" is strictly related to the concept of "*information-as-(an intepretation)*process," such that "there is no knowledge without both an object of knowledge and a knowing subject. The claim that there is absolute knowledge, or knowledge in itself, above and beyond concrete knowing subjects, is fantastic" [23], [24]. Based on this definition, the present work adopts the expression "*information/knowledge processing system*," whose input consists of sensory data and/or information at a lower level of user-specific informative value or utility.
- The rest of this paper considers the following terms as synonyms [16], [17].
  - *Symbolic, semantic, cognitive, categorical, ordinal, nominal, qualitative, subjective, equivocal*. For example, (discrete) categorical variable.
  - *Sub-symbolic, sensory, numerical, non-semantic, quantitative, objective, unequivocal*. For example, continuous or discrete sensory variable.

B. Inductive/Deductive/Hybrid Inference, Either Sub-Symbolic or Symbolic

- There are two classical types of inference (learning) known as *induction*, progressing from particular cases (e.g., true facts, training data samples, etc.) to a general estimated dependency or model, and *deduction*, progressing from a general model to particular cases (e.g., output values) [64].
- *In the words of Mulier and Cherkassky: "induction amounts to forming generalizations from particular true facts. This is an inherently difficult (ill-posed) problem and its solution requires a priori knowledge in addition to data"* [32] (p. 39). That is to say, to become better posed (conditioned) for numerical treatment any (inherently ill-posed) inductive data learning algorithm requires (prior knowledge-based) deductive inference mechanisms to avoid starting from scratch [16].
- The following terms are synonyms of deductive inference and become interchangeable in the rest of this work [16], [17].
  (Sub-symbolic or symbolic) *deductive inference, deductive learning, top-down inference system, coarse-to-fine inference, driven-by-knowledge inference, learning-by-rules, physical model, prior knowledge-based decision system, rule-based system, expert system, syntactic inference, syntactic pattern recognition*.
- The following terms are synonyms of inductive inference and become interchangeable in the rest of this paper [16], [17].
  (Sub-symbolic or symbolic) *inductive inference, inductive learning, bottom-up inference, fine-to-coarse inference, driven-without-knowledge (knowledge-free) inference, learning-from-examples, statistical model*.
- In the rest of this work, expressions like *sub-symbolic (either discrete or continuous) variable*, *symbolic (necessarily discrete) variable*, *sub-symbolic information*, and *symbolic information* are adopted, where sub-symbolic information is a synonym of quantitative data or "*information-as-thing*" while symbolic information is a

synonym of "*information-as-(an interpretation) process*" [23], [24] (refer to Section II-A). Thus, expressions like *inductive/deductive/hybrid (combined deductive and inductive) inference*, either *sub-symbolic* or *symbolic*, are adopted in the rest of this work, depending on whether the inference system deals with, respectively, sub-symbolic continuous/discrete variables or (symbolic and discrete) categorical variables. For example, SIAM™ is a (semi-)symbolic, static (non-adaptive to input data), syntactic (deductive) pre-attentive vision first stage classifier [5]–[17] (refer to Section II-G below).

- In the RS literature, typical examples of sub-symbolic inductive and sub-symbolic deductive inference are, respectively, principal component analysis and tasseled cap transformation [32]–[34].
- In the machine learning literature, typical examples of symbolic inductive inference systems capable of learning from labeled (supervised) data are artificial neural networks, support vector machines, nearest-neighbor classifiers, etc. [32]–[34]. Typical examples of sub-symbolic inductive inference systems capable of learning from unlabeled (unsupervised) data are the unsupervised data clustering algorithms [32]–[34], probability density function estimators, vector data quantizers [32]–[34], image segmentation algorithms [16], [17], [38]–[45], etc.

C. Human Vision

- "A fact that is generally ignored is that biological vision is currently the only measure of the incompleteness of the current stage of computer vision and illustrates that the problem is still open to solution" [47] (refer to Section I). For example, the present paper, which deals with computer vision, namely, RS image understanding (classification) and its quality assessment, should keep human vision as its gold standard.
- The goal of an IUS is to provide plausible (multiple) symbolic description(s) of a 3-D scene, belonging to the (4-D) world-through-time and depicted in a (2-D) image at a given acquisition time, by finding associations between sub-symbolic image features (image-objects or, *vice versa*, image-contours) with symbolic classes of 4-D objects-through-time (4-D concepts-through-time, e.g., buildings, roads, etc.), which belong to a so-called *world model* [35]. Equivalent to a *4-D spatio-temporal ontology of the world-through-time* [25], the world model can be graphically represented as a *semantic network* (*concept network*, traditionally adopted in computer vision, artificial intelligence, machine learning, Geographic Information science (GIS-science), etc.), consisting of: 1) classes of 4-D objects-through-time (concepts) as nodes and 2) inter-concept relations (equivalent to subsets of the Cartesian product between elements of the two concept sets) as arcs between nodes, e.g., spatial relations, either topological (e.g., adjacency, inclusion, etc.) or non-topological (e.g., distance, in-between angle, etc.), non-spatial relations (e.g., part-of, subset-of), or temporal relations [36], [37], [79], [80].
- In mammals, a vision system is comprised of a *pre-attentive vision first phase* and an *attentive vision second phase* summarized as follows.
  (i) Pre-attentive (low-level) vision extracts picture primitives based on general-purpose image processing criteria independent of the scene under analysis. It acts in parallel on the entire image as a rapid (< 50 ms) scanning system to detect variations in simple visual properties [27]–[29]. It is known that the human visual system employs at least four spatial scales of analysis [30], where cells in visual cortex feature gradations of orientation much finer than 45° [144], e.g., around 15° [145]. Single opponent and double opponent color cells are called Type I and Type II, respectively, by Wiesel and Hubel [146] (examples of Type I and Type II receptive fields can be found in [147]). Receptive fields that are spatially opponent, but not color opponent are termed Type III [147].
  (ii) Attentive (high-level) vision operates as a careful scanning system employing a focus of attention mechanism. Scene subsets, corresponding to a narrow aperture of attention, are observed in sequence and each step is examined quickly (20–80 ms) [27]–[29].
- In terms of computational theory, *the problem of image understanding (vision)*, from sub-symbolic (2-D) imagery to symbolic description(s) of the 3-D viewed-scene belonging to the 4-D world-through-time, *belongs to the family of symbolic inductive data learning problems* [16] (refer to Section II-B). As such, it is inherently ill-posed in the Hadamard sense [32] and, consequently, very difficult to solve due to: 1) the well-known *information gap* between varying quantitative sensations (e.g., image features) and stable qualitative percepts (e.g., 3-D object-models belonging to the world model) and 2) the *intrinsic insufficiency of image features* due to occlusion phenomena and dimensionality reduction [16], [35]. Since *vision is an (inherently ill-posed) symbolic inductive inference problem, then its solution requires symbolic prior knowledge in addition to (sub-symbolic) sensory data to become better posed (conditioned) for numerical treatment* [16], [32] (refer to Section II-B). In the literature of psychophysics, according to Vecera and Farah, pre-attentive image segmentation is an interactive (hybrid) inference process "in which top-down knowledge partly guides lower level processing" [31] (p. 1294). This means that *human vision is a symbolic hybrid (combined deductive and inductive) inference system where the ignition of symbolic prior knowledge starts at the pre-attentive vision first stage* [16], [17].

D. QA4EO

Founded in 2003, the GEO is a voluntary partnership of governments and international organizations whose mandate is to provide a framework for the coordination of efforts and strategies capable of addressing common goals in EO disciplines. In 2005, GEO launched a "ten-year implementation plan" to establish the visionary goal of the GEOSS initiative [4].

The GEOSS key objective is to *deliver operational, comprehensive, and timely "knowledge/information products"* (refer to Section I) generated (rather than extracted [23], [24]) from a variety of satellite, airborne, and *in situ* sensory data sources [3]. Interoperability in terms of synergistic use of multi-source multi-resolution data depends upon the successful implementation of two key principles-*Accessibility/Availability* and *Suitability/Reliability*, to allow the provision of and access to *the Right Information, in the Right Format, at the Right Time, to the Right People, to Make the Right Decisions*. This is tantamount to saying that, *according to the GEO* [3], *the necessary and sufficient condition for the development of satellite-based information/knowledge processing systems to be used in operational mode in local- to global-scale monitoring programs is the successful implementation of the GEOSS key objectives of* [4]: *1) Accessibility/Availability and 2) Suitability/Reliability of RS data and data-derived information/knowledge products*.

To pursue the two aforementioned GEOSS key principles, the GEO identified the need to develop a GEO data quality assurance (QA) strategy where *calibration* and *validation (Cal/Val)* activities become critical to data QA and thus to data usability. According to the GEO-CEOS QA4EO guidelines [3]:

(a) An appropriate coordinated program of Cal/Val activities throughout all stages of a spaceborne mission, from sensor building to end-of-life, is considered mandatory to ensure the harmonization and interoperability of multi-source multi-temporal observational data and data-derived products.
(b) To accomplish *validation*, sensory data and -derived products generated in *each step of a satellite-based information processing workflow must have associated with them a set of mutually uncorrelated, quantifiable, metrological/statistically-based QIs featuring a degree of uncertainty in measurement*, to provide a documented traceability of the propagation of errors through the information processing chain in comparison with established community-agreed reference standards (refer to Section I).

By definition, *radiometric calibration* is the transformation of dimensionless digital numbers (DNs) into a community-agreed physical unit of radiometric measure. In line with the QA4EO recommendations, the RS community regards as an indisputable fact that "the prerequisite for physically based, quantitative analysis of airborne and satellite sensor measurements in the optical domain is their calibration to spectral radiance" [81, p. 29]. According to related works [5]–[17], in addition to ensuring the harmonization and interoperability of multi-source observational data, *radiometric calibration is a necessary not sufficient condition for automatic (hybrid model based*, refer to Section II-B) *interpretation of EO imagery*. Irrespective of this common knowledge, radiometric calibration is often neglected in the RS literature and surprisingly ignored by scientists, practitioners, and institutions in RS common practice, including large-scale spaceborne image mosaicking and mapping, e.g., see [82]. For example, in conflict with the QA4EO guidelines, popular RS-IUS commercial software products, such as those listed in Table I, do not consider radiometric calibration of RS imagery as a pre-requisite, with the sole exception of the physical model-based Atmospheric/Topographic Correction (ATCOR-2/3/4) commercial software [83]–[86]. This implies that *popular RS-IUS commercial software products, but the ATCOR-2/3/4, are statistical model-based systems inherently site specific and semi-automatic* [16], [17], [87] (refer to Section II-B).

To be community agreed, a proposed list of QIs of operativeness (OQIs), suitable for the assessment and comparison of RS-IUSs used in operational mode, is summarized below [5]–[17].

(i) Degree of automation (ease-of-use), monotonically decreasing with the number of system free parameters to be user defined. It is also affected by the physical meaning, if any, and the range of variation (e.g., bounded, unbounded, normalized) of the system free parameters.
(ii) Effectiveness or accuracy, e.g., thematic and spatial accuracy of a classification map.
(iii) Efficiency, e.g., computation time and memory occupation.
(iv) Robustness to changes in input parameters.
(v) Robustness to changes in the input data set acquired across time, space, and sensors.
(vi) Scalability, to cope with changes in input data specifications and user requirements.
(vii) Timeliness, defined as the time span between sensory data collection and data-derived product generation. It increases monotonically with computer power and manpower (e.g., the manpower required to collect reference samples for training an inductive data learning system).
(viii) Costs, which increase monotonically with computer power and manpower.

According to the definition promoted by the CEOS WGCV—Land Product Validation (LPV) subgroup, *validation* is the *process of assessing, by independent means, the quantitative accuracy of high-level information products derived from RS data* [2], [48].

In the broader definition promoted by the QA4EO guidelines and adopted in this work, *validation refers to the process of estimating, by independent means, all OQIs, including accuracy as a special case, selected to parameterize a satellite-based information/knowledge processing system for assessment and comparison purposes*.

Extended to the QA4EO definition of validation activities [3], the CEOS hierarchy of validation is the following [48].

- Stage 1 validation. Product QIs and their uncertainties are assessed using a small (typically < 30) set of geographic locations and time periods by comparison with in situ or other suitable reference data.
- Stage 2 validation. Product QIs and their uncertainties are estimated over a significant (globally representative), widely distributed set of geographic locations and multiple time periods and seasons in comparison with reference *in situ* or other suitable independent sources of reference data. Results are published in the peer-reviewed literature.

- Stage 3 validation. Product QIs and their uncertainties are characterized in a statistically robust way via independent measurements over the global (full) range of geographical locations and for all time periods representing global conditions. Results are published in the peer-reviewed literature.
- Stage 4 validation. Validation results for stage 3 are systematically updated when new product versions are released and as the time series expands.

For example, stage 1 validation is involved with a large majority of the existing RS papers where RS-IUSs are assessed and compared. On the contrary, the combination of the present work with related papers [5]–[17] aims at a stage 3 validation of the SIAM™ pre-attentive vision first stage classifier.

E. Probability and Nonprobability Sampling Protocols

- By definition, probability sampling must satisfy the three constraints listed in Section I[55]. Probability sampling methods can be split into equal or variable (unequal) probability sampling methods. Unequal inclusion probabilities create no difficulties as long as they are known and accounted for in the estimation formulas, but equal probability designs possess the advantage of simpler analysis. For example, an area sampling protocol selects polygons into the sample with probability proportional to polygon area, so larger polygons will have a higher probability of being selected [55]. Similar considerations hold for STRS featuring unequal inclusion probabilities. Stratified sampling with proportional allocation results in equal inclusion probabilities, but stratified sampling with either equal or optimal allocation usually leads to different inclusion probabilities for the sampling units in different strata [60]. In general, a different variance estimator formula arises for each accuracy index and each different sampling design [55]. Unlike stated in [88], it is not true that probability sampling is required for assessing the uncertainty of the accuracy estimates.
- Nonprobability sampling methods are all the sampling methods that do not satisfy the requirements of probability sampling methods listed in Section I. According to literature [55]: "unfortunately, examples of nonprobability sampling are common in accuracy assessment applications. Selecting reference locations by purposeful, convenient, or haphazard procedures does not allow the sampling design to determine the inclusion probabilities for each sampling unit. Such designs, therefore, are not probability samples. Purposefully selecting training data for a supervised classification is a good example of a nonprobability sample. Such samples are acceptable for developing a landcover classification map, but often have limited use for accuracy assessment because the necessary probability foundation to permit generalization from the sample data to accuracy of the full population is lacking" [55]. To recapitulate, "it is possible to obtain useful information from nonprobability samples, but the limitations of such data should be recognized." [55]. That said, it is not true that nonprobability sampling methods are unable to provide a degree of uncertainty in measurement, which is contrast with what claimed in [88].
- A *protocol*, defined as a sorted set of guidelines for good practice [48], encompasses a structural and a procedural knowledge (refer to Section I). The definition of international guidelines for best practices, together with standardization, has been a major challenge for the RS community [3], [48], [88]. For example, in this work, the proposed probability sampling protocol complies with community-agreed best practices promoted by the GEO QA4EO guidelines [3] (refer to Section II-D).

F. Geographic Object-Based Image Analysis (GEOBIA)

- In [39], [40], [79], GEOBIA is defined as a sub-discipline of GIScience [25], [141], also known as geomatics engineering [69], devoted to:
  1) The automatic or semi-automatic partitioning (segmentation, aggregation, simplification) of a raster RS image, consisting of sub-symbolic unlabeled pixels, into discrete sub-symbolic labeled image-objects (segments, polygons, regions), where the sub-symbolic label is a segment identifier (e.g., an integer number, say, Segment 1, Segment 2, etc.), such that each discrete image-object is a connected set of pixels whose visual (appearance, pictorial) properties are considered relatively homogeneous with respect to their surroundings according to a measure of similarity chosen subjectively based on its ability to create "interesting" ("meaningful") image-objects [16], [17].
  2) The automatic or semi-automatic mapping (projection) of sub-symbolic labeled image-objects onto a discrete and finite set of LC classes, i.e., of symbolic 4-D object-models-through-time belonging to a *world model* [25], [35], [80], depending on the image-object-specific spatial, spectral, and temporal characteristics, so as to generate as output symbolic vector geospatial information (e.g., LC and LCC maps) in a Geographic Information System (GIS)-ready format.
- About the GEOBIA commitments, Hay and Castilla propose that "the primary objective of GEOBIA as a discipline is to develop appropriate theory, methods and tools sufficient to replicate (and or exceed experienced) human interpretation of RS images in automated/semi-automated ways, that will result in increased repeatability and production, while reducing subjectivity, labor and time costs" [39], [40]. In [79], Lang states that since automation is the overall aim of GEOBIA (like that of any other computer-based technique), the ultimate benchmark of GEOBIA is to mimic human perception.
- Since the year 2000, contemporary with the availability of the first spaceborne VHR commercial images acquired by the GeoEye IKONOS multi-spectral (MS) sensor, two-stage non-iterative GEOBIA systems and three-stage iterative geographic object-oriented image analysis (GEOOIA) systems, where the former is a special case of the latter, i.e., GEOOIA $\supset$ GEOBIA [16], [17], have quickly gained widespread popularity, particularly in Europe, due to the

availability of a series of commercial software products developed by a German company [38], [89]–[92].

- To date a large portion of the RS community considers the GEOBIA/GEOOIA paradigm the state-of-the-art in both scientific and commercial applications of spaceborne/airborne VHR imagery.
- Unfortunately, despite its commercial success, the GEOBIA/GEOOIA approach remains affected by a lack of research, general consensus, and productivity, as acknowledged by increasing sections of the existing literature [16], [17], [38]–[40], [79]. For example, the GEOBIA claim of mimicking human vision [39], [40] remains more an expression of intentions than a fact [16], [17]. To comply with human vision, an artificial vision system should also be a symbolic hybrid inference system in both the pre-attentive vision first stage and the attentive vision second stage (refer to Section II-C). On the contrary, as an example, state-of-the-art GEOBIA/GEOOIA systems share the same inherently ill-posed, driven-without-knowledge, sub-symbolic, inductive image segmentation pre-attentive vision first stage [16], [17], [38]–[45], and feature, at the attentive vision second stage, either an inductive supervised data learning classifier or a symbolic syntactic classifier (also refer to Section II-B). In practice, both GEOBIA and GEOOIA architectures support fully statistical implementations, where no physical model-based inference, equivalent to prior knowledge of the 4-D spatio-temporal world, is ignited.

G. SIAM™

- *The physical model-based (deductive) SIAM™ preliminary classifier is by no means alternative, but complementary in nature to any (inherently ill-posed* [32]*) inductive (statistical model-based) data learning system, either symbolic* (e.g., *artificial neural networks, support vector machines*, etc. [32]–[34]) *or sub-symbolic* (e.g., *unsupervised data clustering algorithms* [32]–[34], *vector data quantization* [32]–[34], *image segmentation algorithms* [16], [17], [38]–[45], etc.), refer to Section II-B [5]–[17].
- To the best of these authors' knowledge, SIAM™ is the first symbolic syntactic inference system (refer to Section II-B), made available to the RS community for operational use in a RS-IUS pre-attentive vision first stage (refer to Section II-C), capable of accomplishing multi-scale image segmentation and multi-granularity image pre-classification simultaneously, automatically and in near real-time [5]–[17].
- In terms of computationl theory (system design, system architecture), exploitation of a pixel-based, symbolic, syntactic pre-attentive vision first stage, like the SIAM™ or the Spectral Classification of surface reflectance signatures (SPECL) implemented as a by-product in the ATCOR-2/3/4 software toolbox [83]–[85], [115] (refer to Table I), referred hereafter as the ATCOR-SPECL sub-system, allows the attentive vision second-stage classification to benefit from driven-by-knowledge regularization of the multiple solution space while avoiding the typical disadvantage of stratification, where identification of informative strata may be difficult [5]–[17]. In other words, SIAM™ should never be considered as a standalone system, but as a module in a three-stage, hierarchical, stratified, feedback RS-IUS architecture consisting of:
  (i) a RS image pre-processing stage zero,
  (ii) a symbolic, syntactic, context-insensitive pre-attentive vision first stage, e.g., implemented as SIAM™ or the ATCOR-SPECL,
  (iii) a battery of attentive vision second-stage, context-sensitive, stratified, feature extractors and one-class classification modules and
  (iv) a feedback mechanism between the pre-attentive vision first stage and the RS image pre-processing stage zero [16], [17].

  Hence, SIAM™ provides this novel hybrid RS-IUS architecture with spectral prior knowledge of the 4-D world-through-time starting from the pre-attentive vision first stage (refer to Section II-C), which makes the inherently ill-posed RS image interpretation problem better posed for numerical treatment (refer to Section II-B).
- The aforementioned three-stage hybrid RS-IUS architecture, employing SIAM™ as its symbolic, syntactic, pre-attentive vision first stage (refer to Section II-C), is alternative to state-of-the-art two-stage non-iterative GEOBIA and three-stage iterative GEOOIA system architectures whose pre-attentive vision first stage consists of an inherently ill-posed sub-symbolic inductive image segmentation algorithm (refer to Section II-F).
- As input SIAM™ requires a RS image radiometrically calibrated into top-of-atmosphere (TOA) reflectance (TOARF) or surface reflectance (SURF) values, where SURF is a special case of TOARF in very clear-sky conditions and flat terrain conditions [93], i.e., TOARF $\supset$ SURF, in compliance with the *Cal(/Val)* requirements of the QA4EO guidelines [3] (refer to Section II-D). In practice, SIAM™ does not consider preliminary atmospheric correction as mandatory (see Table I) because SIAM™ is knowledgeable on how to cope with RS data affected by atmospheric effects (noise). In other words, SIAM™ is capable of recognizing surface types in RS images by "looking through" atmospheric effects, like the presence of haze and thin clouds [5]–[17]. This "look-through" capability is due to the fact that the original prior knowledge base of SIAM™ consists of a reference dictionary of spectral signatures in TOARF values, where relation TOARF $\supset$ SURF means that TOARF $\approx$ SURF $+$ atmospheric noise, whereas traditional libraries of spectral signatures are in SURF values (measured at the ground level), i.e., are atmospheric noise free. Well-known examples of reference dictionaries of spectral signatures in (atmospheric noise-free) SURF values can be found in the existing literature (e.g., refer to [94, p. 273]) or in commercial software products [135], like the U.S. Geological Survey (USGS) mineral and vegetation spectral libraries, the Johns Hopkins University spectral library, and the Jet Propulsion Laboratory mineral spectral library

[83]–[86]. Being provided with an atmospheric noise model, SIAM™ is robust to the presence of atmospheric effects.

- As output SIAM™ delivers preliminary classification (pre-classification) maps at various degrees of semantic granularity. In these semantic maps, the map legend is a discrete and finite set of symbolic informational primitives called *color-based inference categories, spectral-based semi-concepts, spectral categories* or *spectral end members*, e.g., "*vegetation*," "*bare soil or built-up*," "*water or shadow*," etc.
- The semantic meaning of a semi-concept is: 1) superior to zero, which is the semantic value of sub-symbolic *image features*, i.e., image-objects or, *vice versa*, image-contours; and 2) equal or inferior to the semantic meaning of the attentive vision *concepts* (e.g., LC classes, say, "*needle-leaf forest*"), belonging to a *world model*, equivalent to a 4-D spatio-temporal ontology of the physical world-through-time (refer to Section II-C).
- Spectral categories generated as output by SIAM™ belong to six *parent spectral categories* (also called *super-categories*) or major spectral end members which are listed below.

  1) "*Clouds*."
  2) "*Either snow or ice.*"
  3) "*Either water or shadow*."
  4) "*Vegetation*," equivalent to "*either woody vegetation or cropland or grassland (herbaceous vegetation) or (shrub and brush) rangeland*."
  5) "*Either bare soil or built-up*."
  6) "*Outliers*."

  Due to the presence of class "*Outliers*" ("*Unknowns*"), SIAM™ provides a mutually exclusive and totally exhaustive mapping of the input MS image into a discrete and finite set of spectral categories. This is in line with the Congalton and Green requirements of a classification scheme [51]. Although the definition of a rejection rate is a well-known objective of any RS image classification system, e.g., refer to [94], in RS, common practice image classifiers are often applied without any outlier detection strategy.
- Spectral categories in the (2-D) image domain are not LC classes, equivalent to 4-D object-through-time models in the real (4-D) world-through-time model [5]–[17]. In general, one spectral category can belong to many LC classes (e.g., spectral category "*strong vegetation*" can belong to LC classes "*grassland*" or "*agricultural fields*"). Analogously, one LC class encompasses different colors (e.g., class "*deciduous forest* looks like several tones of green equivalent to the quantized colors "*average vegetation*" or "*dark vegetation*"). To conclude, in general, a finite set of many-to-many associations holds between SIAM™'s spectral-based semi-concepts (belonging to the (2-D) image domain) and reference LC classes (equivalent to concepts or 4-D object-models in the real world-through-time).
- SIAM™ is implemented as an integrated system of six sub-systems, including one "master" Landsat-like subsystem plus five "slave" subsystems, whose spectral resolution overlaps with, but is inferior to, Landsat's, refer to Table II.
  1) A "master" seven-band Landsat-like SIAM™ (L-SIAM™) capable of detecting 95/47/18 mutually exclusive and totally exhaustive spectral categories at fine/intermediate/coarse semantic granularity, where symbolic parent–child relationships can be leveraged to improve the RS image interpretation process. The legend of the preliminary classification map generated by L-SIAM™ at fine semantic granularity and consisting of 95 spectral categories is shown in Table III.
  2) A four-band Satellite Pour l'Observation de la Terre (SPOT)-like SIAM™ (S-SIAM™), which detects 68/40/15 mutually exclusive and totally exhaustive spectral categories at fine/intermediate/coarse semantic granularity.
  3) A four-band National Oceanic and Atmospheric Administration Advanced Very High Resolution Radiometer (AVHRR)-like SIAM™ (AV-SIAM™), which detects 82/42/16 mutually exclusive and totally exhaustive spectral categories at fine/intermediate/coarse semantic granularity.
  4) A five-band ENVISAT Advanced Along-Track Scanning Radiometer-like SIAM™ (AA-SIAM™), which detects 82/42/16 mutually exclusive and totally exhaustive spectral categories at fine/intermediate/coarse semantic granularity.
  5) A four-band QuickBird-like SIAM™ (Q-SIAM™), which detects 52/28/12 mutually exclusive and totally exhaustive spectral categories at fine/intermediate/coarse semantic granularity. The legend of the preliminary classification map generated by Q-SIAM™ at fine semantic granularity and consisting of 52 spectral categories is shown in Table IV.
  6) A three-band Disaster Monitoring Constellation-like SIAM™ (D-SIAM™), which detects 52/28/12 mutually exclusive and totally exhaustive spectral categories at fine/intermediate/coarse semantic granularity.

  The output spectral categories detected by the six SIAM™ sub-systems at fine, intermediate, and coarse semantic granularity, described in Table II, are summarized in Table V.
- With regard to implementation, in [6], enough information is provided for the crisp L-SIAM™ implementation to be reproduced. The down-scaled S-SIAM™, AV-SIAM™, and Q-SIAM™ versions generated from L-SIAM™ (refer to Table II) are described in [7], [8]. In [12], the crisp-to-fuzzy SIAM™ transformation is explained in detail. It is noteworthy that since its first 2006 release presented in [6], L-SIAM™ has increased its number of output spectral categories from 46 to 95 (see Table V). This shows that, in line with theory [79], [87], there is a slow "learning curve" in the development and fine-tuning of physical models, like SIAM™.

TABLE II
SIAM™ SYSTEM OF SYSTEMS: LIST OF SPACEBORN/AIRBORNE IMAGING SENSORS ELIGIBLE FOR USE

	Table acronyms. Y: Yes, N: No, C: Complete, I: Incomplete (radiometric calibration offset parameters are set to zero), (E)TM: (Enhanced) Thematic Mapper, B: Blue, G: Green, R: Red, NIR: Near Infra-Red, MIR: Medium Infra-Red, TIR: Thermal Infra-Red, SR: Spatial Resolution, Pan: Panchromatic. **Column highlight color**. Blue columns are related to visible channels typical of water and haze; Green column identify the NIR band, typical of vegetation; Brown columns are related to MIR channels, characteristic of bare soils; Red column: TIR channel, useful to detect fire.											
SIAM™ system of systems		**B – (E)TM1, 0.45-0.52 (μm)**	**G – (E)TM2, 0.52-0.60 (μm)**	**R – (E)TM3, 0.63-0.69 (μm)**	**NIR – (E)TM4, 0.76-0.90 (μm)**	**MIR1 – (E)TM5, 1.55-1.75 (μm)**	**MIR2 – (E)TM7, 2.08-2.35 (μm)**	**TIR – (E)TM6, 10.4-12.5 (μm)**	**SR (m)**	**Rad. Cal. Y/N, C/I**	**Pan SR (m)**	**Notes**
L-SIAM™ (95/47/18 Sp. Cat.)	Landsat-4/-5 TM	×	×	×	×	×	×	×	30	Y-C		Refer to Table I in [6].
	Landsat-7 ETM+	×	×	×	×	×	×	×	30	Y-C	15	Same as above.
	MODIS	×	×	×	×	×	×	×	250, 500, 1000	Y-C		Same as above.
	ASTER		×	×	×	×	×	×	15-30	Y-C		Same as above.
	CBERS-2B	×	×	×	×	×	×	×		N		
	APEX	×	×	×	×	×	×		1.8	Y		Airborne hyperspectral, 285 bands
S-SIAM™ (68/40/15 Sp. Cat.)	SPOT-4 HRVIR		×	×	×	×			20	Y-I	10	Refer to Table II in [6].
	SPOT-5 HRG		×	×	×	×			10	Y-I	2.5 - 5	Same as above.
	SPOT-4/-5 VMI		×	×	×	×			1100	Y-I		Same as above.
	IRS-1C/-1D LISS-III		×	×	×	×			23.5	Y-I		
	IRS_P6 LISS-III		×	×	×	×			23.5	Y-I		
	IRS_P6 AWiFS		×	×	×	×			56	Y-I		
AV-SIAM™ (82/42/16 Sp. Cat.)	NOAA AVHRR			×	×	×		×	1100	Y		Refer to Table II in [6].
	MSG			×	×	×		×	3000	Y		Same as above.
AA-SIAM™ (82/42/16 Sp. Cat.)	ENVISAT AATSR		×	×	×	×		×	1000	Y		Same as above.
	ERS-2 ATSR-2		×	×	×	×		×	1000	Y		
Q-SIAM™ (52/28/12 Sp. Cat.)	IKONOS-2	×	×	×	×				4	Y	1	
	QuickBird-2	×	×	×	×				2.4	Y	0.61	
	GeoEye-1	×	×	×	×				1.64	Y	0.41	
	OrbView-3	×	×	×	×				4	N	1	
	RapidEye-1 to -5	×	×	×	×				6.5	Y-I		
	ALOS AVNIR-2	×	×	×	×				10	Y		
	KOMPSAT-2	×	×	×	×				4	N	1	
	TopSat	×	×	×	×				5	N	2.5	
	FORMOSAT-2	×	×	×	×				8	Y	2	
	ENVISAT MERIS	×	×	×	×				300	Y		Super-spectral, 15 bands
	Leica ADS40/80	×	×	×	×				0.25	Y	0.25	Airborne, 4 bands + PAN
D-SIAM™ (52/28/12 Sp. Cat.)	Landsat-1/-2/-3/-4/-5 MSS		×	×	×				79	Y		
	IRS_P6 LISS-IV		×	×	×				5.8	Y-I		
	SPOT-1/-2/-3 HRV		×	×	×				20	Y-I	10	
	DMC		×	×	×				22-32	Y-C		

TABLE III
PRELIMINARY CLASSIFICATION MAP LEGEND ADOPTED BY L-SIAM™ AT FINE SEMANTIC GRANULARITY CONSISTING OF 95 SPECTRAL CATEGORIES (REFER TO TABLE II). PSEUDO-COLORS OF THE SPECTRAL CATEGORIES ARE GROUPED ON THE BASIS OF THEIR SPECTRAL END MEMBER (E.G., "BARE SOIL OR BUILT-UP") OR PARENT SPECTRAL CATEGORY (E.G., "HIGH" LEAF AREA INDEX (LAI) VEGETATION TYPES). THE PSEUDO-COLOR OF A SPECTRAL CATEGORY IS CHOSEN SO AS TO MIMIC NATURAL COLORS OF PIXELS BELONGING TO THAT SPECTRAL CATEGORY

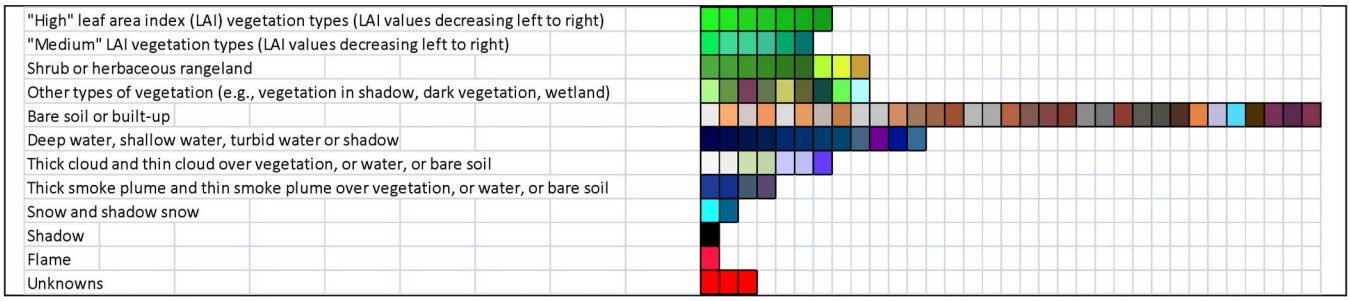


TABLE IV
PRELIMINARY CLASSIFICATION MAP LEGEND ADOPTED BY Q-SIAM™ AT FINE SEMANTIC GRANULARITY CONSISTING OF 52 SPECTRAL CATEGORIES (REFER TO TABLE II). PSEUDO-COLORS OF THE SPECTRAL CATEGORIES ARE GROUPED ON THE BASIS OF THEIR SPECTRAL END MEMBER (E.G., "BARE SOIL OR BUILT-UP") OR PARENT SPECTRAL CATEGORY (E.G., "HIGH" LEAF AREA INDEX (LAI) VEGETATION TYPES). THE PSEUDO-COLOR OF A SPECTRAL CATEGORY IS CHOSEN SO AS TO MIMIC NATURAL COLORS OF PIXELS BELONGING TO THAT SPECTRAL CATEGORY

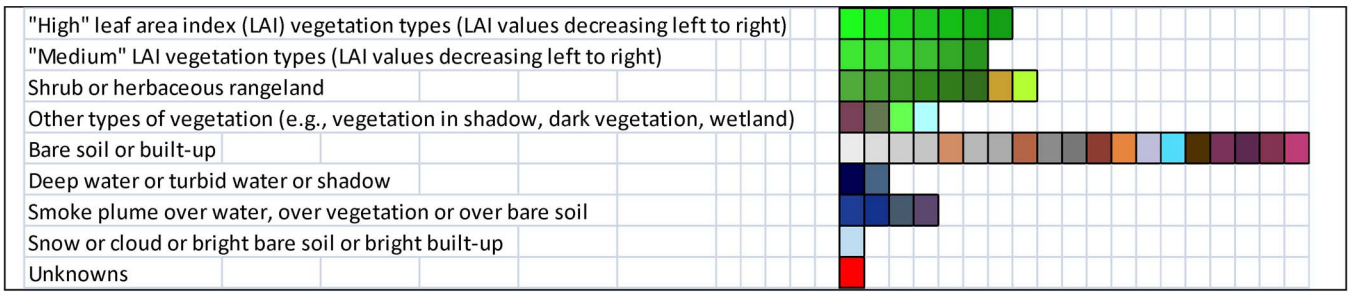


TABLE V
SIAM™ SYSTEM OF SYSTEMS. SUMMARY OF INPUT BANDS AND OUTPUT SPECTRAL CATEGORIES REPORTED IN TABLE II. (∗) EMPLOYED IN SENSOR-INDEPENDENT BI-TEMPORAL POST-CLASSIFICATION CHANGE DETECTION

SIAM™	**Input Bands (B: Blue, G: Green, R: Red, NIR: Near Infra-Red, MIR: Medium IR, TIR: Thermal IR)**	**Preliminary Classification Map Output Products: Number of Output Spectral Categories.**			
		Fine Semantic Granularity	**Intermediate Semantic Granularity**	**Coarse Semantic Granularity**	**Inter-Sensor Semantic Granularity (*)**
L-SIAM™	7 – B, G, R, NIR, MIR1, MIR2, TIR	95	47	18	33
S-SIAM™	4 – G, R, NIR, MIR1	68	40	15	
AV-SIAM™	4 – R, NIR, MIR1, TIR	82	42	16	
AA-SIAM™	5 – G, R, NIR, MIR1, TIR	82	42	16	
Q-SIAM™	4 – B, G, R, NIR	52	28	12	
D-SIAM™	3 – G, R, NIR	52	28	12	

- In terms of OQIs (refer to Section II-D), *existing per-pixel physical model-based decision-tree preliminary classifiers, like SIAM™ or the ATCOR-SPECL, detect output spectral categories (symbolic strata, symbolic masks) automatically and in near real-time, where automation does not come at the expense of accuracy or robustness to changes in the input data set, but at the expense of the informative content of spectral categories generated as output information primitives whose semantic value is low*, i.e., *equal or inferior to that of target 4-D LC classes-through-time, refer to Section II-C*.

III. PROBLEM RECOGNITION AND OPPORTUNITY IDENTIFICATION

To pursue the two GEOSS key principles of *Accessibility/Availability and Suitability/Reliability of RS data and data-derived information/knowledge products*, considered necessary to allow the provision of and access to *the Right Information, in the Right Format, at the Right Time, to the Right People, to Make the Right Decisions*, the GEO-CEOS QA4EO guidelines require *validation* of sensory data and data-derived products in terms of *quantifiable, metrological/statistically based QIs*

featuring a degree of uncertainty in measurement [3] (refer to Section II-D).

In recent years, the GEOSS visionary goal of providing harmonized multi-source EO data, data-derived geospatial information products and operational (turnkey, ready-to-go, good-to-go) services at global, regional, and local spatial scales has become increasingly urgent due to multiple drivers. First, cost-free access to large-scale low spatial resolution (LR) (above 40 m) and medium spatial resolution (MR, from 40 to 20 m) spaceborne image databases has become a reality in line with the GEO vision [3], [95]. Second, the demand for HR and VHR commercial satellite imagery has continued to increase in terms of data quantity and quality [1] (refer to Section I). Third, an increasing number of ongoing international research projects aims at delivering operational RS-IUS products and services at global spatial scale [2]. Among these ongoing programs worth mentioning is the Global Monitoring for the Environment and Security (GMES), an initiative led by the European Union (EU) in partnership with the European Space Agency [96], [97], the National Aeronautics and Space Administration (NASA) Land Cover and Land Use Change (LCLUC) program [2, p. 3] and the USGS-NASA Web-Enabled Landsat Data (WELD) project [98], in addition to the aforementioned GEO GEOSS [3], [95].

Unfortunately, to date, the automatic or semi-automatic transformation of huge amounts of multi-source multi-resolution EO images into information/knowledge can still be considered far more problematic than might be reasonably expected (refer to Section I). In practice, the increasing rate of collection of EO data of enhanced spatial, spectral, and temporal quality outpaces the ability of existing RS-IUSs to generate information/knowledge (e.g., LC and LCC maps, also refer to Section II-A) from RS data [5]–[17].

Collected from the existing literature, many kinds of converging evidence support the conclusion that productivity in terms of quality, quantity, and value of RS data-derived products delivered by the RS community can still be considered low, in contrast with the visionary goal of the GEOSS project and the QA4EO guidelines. These converging sources of evidence include the following:

- According to philosophical hermeneutics, the impact upon computer science, information technology, artificial intelligence, and machine learning of existing different quantitative and qualitative concepts of information (namely, "*information-as-thing*" and "*information-as- (an intepretation) process*"), embedded in more or less explicit information theories (refer to Section II-A), appears largely underestimated [23], [24]. It means that fundamental questions-like: When do sub-symbolic data become symbolic information? When does vision go symbolic [21]? etc.—appear largely underestimated and, as a consequence, far from being answered [16], [17].
- There is an ongoing multi-disciplinary debate about a claimed inadequacy of scientific disciplines such as computer vision, artificial intelligence/machine intelligence and cybernetics/machine learning from data, whose origins date back to the late 1950s, in the provision of operational solutions to their ambitious cognitive objectives [18], [19]. Deductive inference is the main focus of interest of traditional artificial intelligence. Inductive inference is the basis of the machine learning discipline. It may mean that, if they are not combined, deductive and inductive inference systems (refer to Section II-B) show intrinsic weaknesses in operating mode, irrespective of implementation [16], [17].
- "Research and development in computer vision is often considered quite separate from research into the functioning of human vision. A fact that is generally ignored is that biological vision is currently the only measure of the incompleteness of the current stage of computer vision, and illustrates that the problem is still open to solution" [47] (refer to Section II-C).
- Typically adopted in LR (coarser than 20 m) RS image applications, traditional context-insensitive (pixel-based) deductive (physical model-based) or inductive (statistical model-based) RS-IUSs, exploiting the sole context-insensitive chromatic/achromatic information in a (2-D) image domain, tend to be affected by a well-known salt-and-pepper classification noise effect [32]–[34]. To outperform traditional pixel-based classifiers, starting from the late '80s when spaceborne HR images (e.g., SPOT-1 to SPOT-5 imagery) became available for scientific and commercial applications, a new generation of context-sensitive RS-IUSs, capable of dealing with image-objects, including (0-D) points, (1-D) lines, (2-D) polygons, and multi-part polygons (strata) according to the Open Geospatial Consortium Simple Feature Specification [33], rather than pixels alone, has been proposed to the RS community [5]–[17], [77], [99], [100]. Since the year 2000, contemporary with the availability of the first spaceborne VHR commercial images (namely, the IKONOS images acquired and distributed by GeoEye), two-stage non-iterative GEOBIA systems and three-stage iterative GEOOIA systems, where the former is a special case of the latter, i.e., GEOOIA $\supset$ GEOBIA [16], [17], have quickly gained widespread popularity [38], [89]–[92]. Unfortunately, despite its commercial success, the GEOBIA/GEOOIA approach remains affected by a lack of research, general consensus, and productivity, as acknowledged by increasing sections of the existing literature [16], [17], [38]–[40], [79] (refer to Section II-F).
- To outperform existing deductive and inductive inference systems, a novel trend in recent literature aims at developing hybrid inference systems capable of continuous and categorical variables extraction from sensory data [35] (refer to Section II-B). For example, to be considered inspired to human vision, an artificial vision system should be implemented as a symbolic hybrid inference system comprising a symbolic hybrid pre-attentive vision first stage (refer to Section II-C). In line with this trend, new opportunities in the design and implementation of operational hybrid RS-IUSs have been proposed to the RS community in recent years [5]–[17]. For example, a three-stage hybrid RS-IUS architecture, employing the operational SIAM™ software product as its symbolic, syntactic, pre-attentive vision first stage (refer to Section II-G), is

proposed as a viable alternative to existing state-of-the-art two-stage non-iterative GEOBIA and three-stage iterative GEOOIA systems, whose pre-attentive vision first stage consists of an inherently ill-posed sub-symbolic inductive image segmentation algorithm (refer to Section II-F).

- Publication standards typically adopted by the RS literature may be considered inadequate to the assessment of alternative RS-IUSs in operating mode.
  - In nonprobability sampling, sampling units are selected by a purposeful, convenient, or haphazard procedure that does not allow to determine the inclusion probability for each sampling unit. Hence, nonprobability sampling lacks the necessary probability foundation to permit generalization from the sample data to the full target population [55] (refer to Section II-E). Whereas nonprobability sampling is perfectly acceptable for, say, training/testing any inductive data learning classifier, it should never be employed for map accuracy validation. On the contrary, in the RS literature, there is a lack of probability sampling protocols enforced for RS data-derived product validation in compliance with principles of statistics and the QA4EO guidelines (as negative examples not to be imitated, refer to [54], [61], [62]).
  - The sole accuracy is typically selected from the possible set of mutually independent OQIs eligible for parameterizing RS-IUSs for assessment and comparison purposes (refer to Section II-D). As a consequence of this experimental drawback, the operational domain of applications of these RS-IUSs remains unknown or appears questionable. For example, how does accuracy of a RS-IUS tested in a local mapping problem scale to regional, continental, global mapping applications [2]?
  - Map accuracy estimates are almost never provided with a degree of uncertainty in measurement in compliance with the principles of statistics together with the QA4EO recommendations [3]. The practical consequence of this experimental drawback is that the statistical significance of these accuracy estimates remains unknown.
  - In general, alternative RS data mapping solutions are tested in toy problems at small spatial scale (e.g., local scale) and/or coarse semantic granularity, i.e., the CEOS WGCV LPV Stage 1 validation requirements tend to be accomplished at best (refer to Section II-D). The practical consequence of this experimental drawback is that the robustness of these RS-IUSs to changes in the input data set together with their scalability to real-world RS applications at large (e.g., continental, global) spatial scale and fine semantic granularity remain unknown or appear questionable [16], [17].
  - It should be well known that first-stage pre-attentive vision algorithms, including image-object segmentation and image-contour detection approaches (the latter being the dual problem of the former), are inherently ill-posed problem in the Hadamard sense [16], [17], [32], [35], [38]–[45] (refer to Section II-C). Irrespective of this scientific evidence, dozens of presumably "better" image segmentation/contour detection methods are presented each year in the RS and computer vision literature, while relatively little research has focused on the development of image segmentation probability sampling strategies for quality assessment and comparison purposes [61], [62], [101], [102]. This lack makes it hard to compare different image segmentation/contour detection methods or even different parameterizations of a single method.
  - Quality assessment protocols or guidelines, like those proposed in [54] and [88], do not satisfy the protocol definition provided in Section II-E. For example, the so-called protocol for accuracy assessment of classification maps generated from VHR images proposed in [54] does not provide a set of rules for accuracy assessment starting from (probability) sampling design to end up with sample analysis and estimation (refer to Section I). In practice, [54] presents a mere list of formulas of thematic accuracy indices and geometric error indices.

As an example of the theoretical and methodological limitations mentioned above, let us consider works on image segmentation assessment based on a finite reference sample set, like [54], [61], and [62], which differ from works focused on a complete-coverage segmentation map pair comparison, such as [142]. In papers like [54], [61], and [62], segmentation QIs are collected against, respectively, 11, one (!), and 37 reference image-objects selected by a nonprobability sampling strategy in a spaceborne VHR image. Hence, they lack the necessary probability foundation to permit generalization from the sample data to accuracy of the whole population [88] (refer to Section II-E). In addition, estimated QIs are not provided with any degree of uncertainty in measurement. Hence, they have no statistical significance [3]. If computed, due to the tiny or small cardinality of the reference sample set, the degree of uncertainty of these QI estimates would be extremely large [50], showing that little (useful) information is conveyed by these QI values.

Based on the aforementioned considerations, it is possible to conclude that, almost ten years from the GEOSS launch, the GEO-CEOS QA4EO guidelines have been successful in gaining attention of the RS community on the GEOSS principle of *Accessibility/Availability* of sensory data and data-derived products. On the other hand, the second GEOSS principle of *Suitability/Reliability* of operational, comprehensive and timely "knowledge/information products" derived from RS data can still be considered far from being accomplished by the RS community.

According to philosophical hermeneutics, the cause of this dichotomy is well known [23], [24]. The first GEOSS key principle is quantitative (unequivocal) and related to the Shannon concept of "*information-as-thing*" irrespective of its meaning [78]. As such, it is easier to deal with than the second GEOSS principle, which is qualitative (equivocal), has to deal with the

meaning (interpretation, understanding) of (quantitative) data, and is related to the concept of "*information-as-(an interpretation) process*" (refer to Section II-A).

To have a favorable impact on the yet-unaccomplished second GEOSS key principle, where *Suitability/Reliability* of operational, comprehensive, and timely RS data-derived information products and services is required, this work aims at two objectives. The main objective is to present to the RS community a novel and, to the best of these authors' knowledge, the first probability sampling protocol for accuracy assessment of thematic maps generated from VHR images in compliance with the QA4EO guidelines (refer to Section I). This means that, among the OQIs listed in Section II-D, this work focuses on the sole mapping accuracy assessment, required to be statistically consistent and statistically significant, in contrast with a major portion of the RS literature where non-probability sampling methods are adopted instead (refer to this section above).

In the experimental session, the proposed protocol is tested in the accuracy validation of thematic maps automatically generated from a test set of VHR images, acquired across time, space, and sensors, by the existing SIAM™ software product (refer to Section II-G). Hence, as its secondary objective, this work provides a statistically consistent and statistically significant accuracy validation of the SIAM™ software product (refer to Section I), eligible for use as the pre-attentive vision first stage in a novel generation of automatic three-stage hybrid RS-IUSs (refer to Section II-G).

To accomplish its primary objective, this work finds several opportunities in the existing literature as described below.

- In [55], Stehman and Czaplewski discuss the fundamental principles of the six basic components of a probability sampling protocol for thematic map accuracy assessment (refer to Section I).
- Overton and Stehman provide a helpful discussion of the Horvitz–Thompson theorem as a unifying perspective for probability sampling [60].
- It is common knowledge that QI selection has to account for the well-known non-injective property of QIs [5]–[17]. This implies that no hypothetical universal QI can exist, which contradicts a significant portion of the RS literature [57]–[59] (refer to Section I). For example, Stehman states that "numerous accuracy measures have been proposed for summarizing the information contained in an error matrix. No one measure is universally best for all accuracy assessment objectives, and different accuracy measures may lead to conflicting conclusions because the measures do not represent accuracy in the same way. Choosing appropriate accuracy measures that address objectives of the mapping project is critical" [49].
- Although often forgotten in RS common practice, it is well known that the spatial distribution of mapping errors, also known as *locational accuracy* [53] or *location error* [56], is not investigated by traditional TQIs [49], [50]–[53], which are typically site insensitive (nonsite specific [53], context-independent), i.e., pixel based. As a consequence, sub-symbolic object-specific (site-specific) SQIs have been proposed [52], [53], [56], to be estimated in combination with the more "traditional" symbolic pixel-based TQIs (refer to Section I). In practice, in agreement with [54], a (symbolic) classification map can be transformed into a (sub-symbolic) segmentation map (refer to footnote 1). Thus, SQIs compare pairs of test and reference polygons (2-D segments) in terms of shape, irrespective of their thematic labels. The complementary nature of symbolic pixel-based TQIs and sub-symbolic object-based SQIs is analogous to the exploitation of both pixels and image-objects, considered complementary rather than alternative, in RS-IUSs such as those proposed in [5]–[17], [77], [99]. Intuitively, sub-symbolic object-based SQIs can be related to the pre-attentive vision first phase in both human vision (refer to Section II-C) and GEOBIA/GEOOIA systems (refer to Section II-F), e.g., refer to [54], [62], [101]. On the other hand, symbolic TQIs can be related to the attentive vision second phase in both human vision (refer to Section II-C) and GEOBIA/GEOOIA systems (refer to Section II-F).
- Optimized mutually uncorrelated symbolic pixel-based TQIs can be selected in compliance with the works by Stehman [49], Foody [53], [103] and Pontius *et al.* [127]. Unfortunately, to date, TQIs promoted by, say, Stehman [49] and Pontius *et al.* [127] are in contrast with a large portion of the existing RS literature where the kappa coefficient of agreement, the zeta significance of the difference in accuracy between two maps with independent kappa coefficients, and the normalization of an error matrix are still very popular, despite their well-known drawbacks [49], [53], [103], [127].
- Mutually uncorrelated sub-symbolic image-object-based SQIs can be inspired to those estimated in [54].
- Variance estimation formulas for both TQI and SQI ensembles can be selected from [50] (refer to Section II-D).
- A general rule of thumb would require to select the reference data source one step closer to the ground than the RS data used to make up the map [51] (refer to Section I). Unfortunately, when dealing with thematic maps generated from VHR imagery, it is often the case there is no reference data source originated at the same time of the VHR image acquisition, but one step closer to the ground. For example, to assess the accuracy of thematic maps generated from, say, the test VHR image set acquired in the year 2010 and adopted in this work (refer to Section IV below), pre-existing VHR thematic maps dated 2010 would be required, since ground visits cannot be performed back in time. In general, in these cases the sole data source available for reference population sampling is the same VHR image adopted as input by the RS-IUS whose output map has to be evaluated. In other words, the test and reference data sources coincide with the VHR image at hand. In compliance with the (qualitative, equivocal) concept of "*information-as-(an intepretation) process*" [23], [24] (refer to Section II-A), the lack of a reference data source one step closer to the ground than the VHR image at hand should not be considered a problem, as far as the second knowledge expert (reference cognitive agent), the one in charge of implementing the sample evaluation and

labeling phases of the map accuracy assessment protocol (refer to Section VI below), interprets the VHR image by independent means from the first (test) cognitive agent, namely, the RS-IUS whose maps are being validated. In general, the collection of reference ("truth") samples, from the photointerpretation of VHR imagery, ground visits, existing maps, tabular data, or a combination of these sources, remains an equivocal ("*information-as-(an intepretation) process*" [23], [24], refer to Section II-A), expensive, tedious, difficult or impossible task [3], [55].

IV. Test Images and Thematic Maps

Since the topic of this work is the accuracy assessment of maps generated from VHR images in a satellite-based information/knowledge processing system workflow in compliance with the QA4EO guidelines [3], the problem of quality assessment of data-derived products should not be considered independent of the radiometric and geometric quality of the input data source (since "garbage in means garbage out").

In the framework of the 2011 DigitalGlobe eight-band Challenge, two eight-band 2 m-resolution off-nadir WorldView-2 (WV-2) images of the capital site of Brazilia (Brazil) acquired in, respectively, the green season (identified as acquisition time 1, Time-1 (T1)) and the dry season (identified as acquisition time 2, Time-2 (T2)) of the year 2010, were provided to the present authors by DigitalGlobe for testing purposes [3]. An additional four-band 2.4-m resolution QuickBird-2 (QB-2) image of Brazilia, acquired in 2010 at time $T1 + 45$ days, was provided by DigitalGlobe for comparison purposes.

In this experimental session, the two WV-2 images and the QB-2 image are radiometrically calibrated into TOARF values, in compliance with the *Cal/Val* requirements of the QA4EO guidelines (see Section II-D) and with the input data requirements of the SIAM™ software toolbox (refer to Section II-G). Second, the two "slave" WV-2 images calibrated into TOARF values are radiometrically registered (re-calibrated) to match two "master" 7-band 30 m-resolution nadir-viewing Landsat-7 Enhanced Thematic Mapper+ (ETM+) images radiometrically calibrated into TOARF values. Next, the three test VHR images, featuring their most advanced calibration (or re-calibration) stage, are mapped automatically by SIAM™. Finally, an automatic post-classification change detection software module is applied to the two SIAM™ maps generated from the re-calibrated WV-2 image pair.

A. WV-2 and QB-2 Image Pre-Processing

RS image pre-processing (enhancement), whose goal is to transform an input image into an output image of enhanced geometric and/or radiometric quality, is clearly acknowledged as a fundamental pre-requisite of RS image quantitative analysis [3], [81], [83]–[87]. The rest of this section focuses exclusively on the radiometric quality enhancement of spaceborne/airborne VHR imagery, i.e., it does not deal with the improvement of the geometric quality of RS images through geo-projeEnhanced Thematic Mapperction, co-registration, and orthorectification, considered beyond the scope of this paper although their importance is fundamental, particularly in multi-temporal analysis, like LCC detection, where co-registration quality is typically required below 1 pixel [83], [84], [113], [114].

It is well known that pictorial properties (reflected radiances) of the same cover type are "affected significantly by Sun-target-sensor geometry because most types of natural and artificial surfaces are anisotropic reflectors... As a result, a fixed target may be viewed from very different viewing angles, leading to varying reflected radiance. This affects the detectability of the temporal evolution of a fixed target. The bidirectional dependency also poses problems in the retrieval of surface parameters such as spectral albedo and radiant fluxes, as they are defined over all directions" [104]. A typical illumination effect is when, caused by self-occlusion or occlusion phenomena [10], more shadows are seen looking toward the sun whereas more illuminated surfaces are seen when looking away from the sun. Another typical illumination effect is when multi-temporal images of the same surface area featuring increased solar zenith angles appear darker and affected by longer shadows [104].

Starting from the late '70s, the non-Lambertian nature of vegetation reflectance has been considered a well-known limitation to the use of AVHRR visible (VIS) and near-infrared (NIR) reflectances. For example, in [105], it is stated that "it would not be possible to composite reflectance data reliably from AVHRR data without LC class-specific bidirectional reflectance distribution function (BRDF) effects correction." To normalize varying AVHRR image acquisition conditions, a time series of AVHRR images acquired at different times and Sun-target-sensor geometries are required as input to fit a model of vegetated BRDF effects; next, vegetation spectral reflectances are standardized to a chosen geometry of view and solar position (e.g., nadir view and 45° solar zenith angle [104], [105]) to exhibit much smoother seasonal variations [104] and smaller inherent data spread [105] than the original reflectances upon fixed vegetation targets, which facilitates statistical analysis and interpretation of RS images for monitoring changes associated with surface conditions.

Two main approaches for correction of radiometric variations in multi-date imagery emerge from the literature: physical and statistical approaches for radiometric normalization [104], [106]. Atmospheric radiative transfer models aim at accounting for some or all radiometric effects on surface response by converting DNs into physical units of TOA radiance (TOARD) using sensor calibration coefficients, then calculating SURF values as a function of sunlight (direct) illumination conditions (e.g., affected by the Sun-to-Earth distance at the acquisition time), topographic effects (e.g., incident angle between the sunlight direction and the normal to the target surface), and atmospheric effects (e.g., airlight component), where the (isotropic) Lambertian surface hypothesis holds. Finally, (anisotropic) non-Lambertian surface types are accounted for by estimating surface albedo from SURF values divided by a LC class-specific BRDF factor [83], [84], [86]. With regard to physically based BRDF models that require accurate preliminary correction of TOARD values into SURF values, e.g., refer to the Second Simulation of the Satellite Signal in the Solar Spectrum (6S) code developed by Vermote *et al.* [107], these are functions (e.g., linear functions) of the LC type,

electromagnetic wavelength, solar, and sensor positions [105]. LC class-specific BRDF models require estimation of model free parameters to be optimized in a set of RS images, acquired in a variety of solar and sensor positions, sufficient to enable BRDF models to be fitted [104]–[106]. In practice, a reference library of sensor-specific multiple views of a fixed target surface type must be built up to fit a LC type-specific BRDF model. This approach typically presents two problems. "First, there is no guarantee that the BRDF of, say, vegetation remains constant over time, for reflectance changes as plants grow and respond to their environment. Second, atmospheric conditions can vary considerably and, unless these are taken into account, will affect the calculated BRDF. Modeling bidirectional reflectance of a variety of vegetation types and performing BRDF correction to a known accuracy remains a difficult task" [105].

The statistical model-based approaches for radiometric correction aim at an empirical radiometric normalization of relative radiometric differences existing between multiple data sets through radiometric registration (matching), i.e., registering one data set to another so that it appears as if both were acquired under the same set of acquisition conditions. Mostly adopted for correction of time series of RS images rather than spatial matching of images, statistical approaches have also been used to generate radiometrically matched image mosaics [106].

This subsection focuses on the radiometric quality of RS images in terms of:

- Absolute radiometric calibration of DNs into TOARD values (refer to Section II-D) [5]–[17].
- Atmospheric correction and topographic correction (TOC) [10], [108], [109] of TOARD into SURF values when the target surface is assumed to be Lambertian [93].
- Approximation of SURF with TOARF values in clear-sky and flat terrain conditions (refer to Section II-D) [5].
- Surface type-specific BRDF effect correction of SURF or TOARF values into spectral albedo values to account for non-Lambertian surfaces [83], [84], [86], [105].

1) Absolute Radiometric Calibration of DNs into TOARD Values: The linear conversion of DNs into radiometrically corrected TOARD values, called *absolute radiometric calibration*, assumes the viewed earth surface be Lambertian (isotropic). For example, for each pixel p and wavelength λ in band *Band* of a WV-2 or QB-2 product acquired at time t[110], like the VHR test images adopted in this work (refer to the introduction to Section IV), the following absolute radiometric calibration equation holds:

$$0 \leq TOARD(p, \lambda \equiv Band, t) = \frac{[absCalFactor(Band,t)^{*}DN(p,Band,t)]}{effectiveBandwidt(Band,t)} \quad (1)$$

where $TOARD(p,\lambda) \geq 0$ is expressed in radiometric physical units $[\mathrm{W} * \mathrm{m}^{-2} * \mathrm{sr}^{-1} * \mathrm{micrometer}^{-1}]$, while absolute calibration parameters *absCalFactor* and *effectiveBandwidth* are found in the WV-2 and QB-2 image medatadata file (extension.IMD).

2) Atmospheric Correction and Topographic Correction of TOARD into SURF Values: In the words of Schaepman-Strub *et al.* [81], "reflectance quantities acquired under hemispherical illumination conditions (i.e., all outdoor measurements) depend not only on the scattering properties of the observed surface, but as well on atmospheric conditions, the object's surroundings, and the topography, with distinct expression of these effects in different wavelengths." For example, the calibrated at-sensor radiance in pixel p of an EO image, where p is located in (*lat, long*) coordinates, at acquisition time t and wavelength λ, identified as $TOARD(p,\lambda,t)$, can be described as [10]

$$0 \leq TOARD(p,\lambda,t) = L_{surface}(p,\lambda,t) + L_{adj}(p,\lambda,t) + L_{airlight}(\lambda,t) \quad (2)$$

where $L_{surface}(p,\lambda,t)$ is the target surface-reflected radiance due to several components, namely, an unscattered *sunlight* component, a scattered *skylight* component [111], and a scattered terrain irradiance incident on the target directly from surrounding terrain slopes (refer to this section below), $L_{adj}(p,\lambda,t)$ is the adjacency radiance reflected by objects other than the target and scattered or reflected into the sensor by the atmosphere, and $L_{airlight}(\lambda,t)$ is the scattered solar radiance from the atmosphere to the sensor, commonly called *airlight* [111] or *upwelling path radiance* [83], [84]. It is noteworthy that terms $L_{adj}(p,\lambda,t)$ and $L_{airlight}(\lambda,t)$ contain no information on the surface properties of the target pixel p.

Expanding term $L_{surface}(p,\lambda,t)$ when the Lambertian surface assumption holds, we obtain [83]–[86]

$$0 \leq L_{surface}(p,\lambda,t) = SURF(p,\lambda,t)\cdot\tau_{uw}(\lambda,t)\cdot\frac{1}{\pi\cdot d_{SE}(t)^2} \cdot[(\tau_{dw}(\lambda,t)\cdot ESUN(\lambda)\cdot\cos(\varphi(p,t))) + E_{dif}(\lambda,t)\cdot s + E_{ter}(p,\lambda,t)] \quad (3)$$

where $SURF(p,\lambda,t) \in [0,1]$ is the target surface reflectance coefficient, $\tau_{\mathrm{uw}}(\lambda,t) \in [0,1]$ is the upward atmospheric spectral transmittance, $\tau_{\mathrm{dw}}(\lambda,t) \in [0,1]$ is the downward atmospheric spectral transmittance, $ESUN(\lambda)$ is the exoatmospheric solar irradiance, found in literature [5] and related to the so-called *sunlight* [111], $\varphi(p,t) \in [0,\pi/2]$ is the incident angle on a tilted surface, where $\varphi(p,t)$ is computed from a digital surface model (DSM) and/or a digital terrain model (DTM) and the sun position, namely, the solar zenith and azimuth angles found in the RS image metadata, $d_{SE}(t)$ is the Earth–Sun distance in astronomical units to be interpolated from values found in literature as a function of the viewing day and time, t, transformed into a Julian day value in range {1, 365}, such that $d_{SE}(t)$ approximately belongs to range $1 \pm 3.5\%$ [5], $E_{\mathrm{dif}}(\lambda,t) \geq 0$ is the hemispherical diffuse irradiance, also called *diffuse irradiance at the surface, ambient light or indirect illumination* [5], s is the *skyview factor* (visible portion of the sky, related to the so-called *skylight*) $\in$ [0, 1] to be computed from a DSM and/or a DTM [83]–[86], and $E_{ter}(p,\lambda,t) \geq 0$ is the scattered/reflected *terrain irradiance* incident on the target directly from surrounding terrain slopes, then this *terrain irradiance* component is null in flat terrains [5].

The substitution of (2) in (3) provides the solution of the simplified radiative transfer equation in terms of surface reflectance, $SURF(p,\lambda,t) \in [0,1]$, where the Lambertian surface assumption holds [83]–[86] (see (4), shown at the bottom of the page).

In (4), atmospheric effects are modeled by atmospheric parameters $\tau_{uw}(\lambda,t) \in [0,1]$, $\tau_{dw}(\lambda,t) \in [0,1] L_{airlight}\ (\lambda,t) \geq 0$ and $E_{dif}(\lambda,t) \geq 0$. Topographic effects, requiring a DEM to be assessed, are the incident angle to the target surface element $\varphi(p,t)$, $E_{ter}(p,\lambda,t)$ and $L_{adj}(p,\lambda,t)$ and the skyview factor s. It is noteworthy that term $L_{adj}(p,\lambda,t)$ accounts for both atmospheric and topographic effects.

With regard to atmospheric effects, in [112], a preliminary study of atmospheric stability was made on a dried lake in Tuz Golu, Turkey, selected as a CEOS vicarious campaign site offering good overall spectral uniformity and ease of access, in August 2010. According to these authors, "data show larger than expected variation in both the day to day and within day measurements. Higher values of the atmospheric aerosol optical thickness (AOT) were collected during the earlier part of campaign interval than expected ranging from 0.5 to 0.65 for the shorter 340 nm wavelengths. Also, it shows substantial changes in AOT halfway through the campaign, dropping to between 0.15 and 0.3. The impacts of having an AOT that ranges from 0.65 to 0.15 for 380 nm is the transmittance of the atmosphere ranges from 0.52 to 0.86. This clearly demonstrates the need for coincident measurements of the atmosphere at the time of overpass. In comparison, a site used in Brookings, South Dakota, during the entire month of August, varies from 0.4 to 0.1, and has an average of 0.15 at 380 nm. This results in a change in the transmittance of the atmosphere at 380 nm between 0.67 and 0.90. Another site near Algodones Dunes in southwestern part of the USA has an AOT at 380 nm varying from 0.2 to 0.05, or a transmittance value of 0.82 to 0.95."

With regard to topographic effects, the well-known problem of RS image TOC is a circular (chicken-and-egg) dilemma: while image classification should be run only after TOC takes place, TOC requires *a priori* knowledge of surface roughness which is LC class specific [10], [108], [109]. To overcome this limitation, "more research regarding the use of better stratification methods" is strongly encouraged [108], [109]. These recommendations are accounted for in [10], where an automatic symbolic stratification of spaceborne optical imagery is accomplished for an automatic TOC implementation via the SIAM™ software toolbox [5]–[17].

In addition to categorical stratification, TOC requires DSM data be subjected to quality constraints. In [113], results reveal that the accuracy of the TOC depends on the accuracy and spatial resolution of the DSM data as well as the co-registration between the DSM and satellite images. In practice: 1) artifacts (e.g., information holes) in the DSM data can cause significant local errors in the correction, 2) mis-registration error of one or two pixels can lead to large error of retrieved surface reflectance values, and 3) a DSM resolution equal or below the spatial resolution of satellite imagery is needed for the best results.

To recapitulate, in TOC applications the following DSM requirements specification holds:

- In line with the RS image orthorectification requirements [114], the DSM spatial resolution should be $\leq (1/4) \div 1$ times the spatial resolution of the imaging sensor [83]–[86], [113], [114].
- The mis-registration error between DSM and orthorectified satellite imagery should be below 1 pixel [113], [114], in line with co-registration requirements for RS image orthorectification [83], [84].
- The DSM quality should be "high" to avoid DSM artifacts (e.g., no holes in the DSM data) [113].

3) Approximation of SURF with TOARF Values in Clear-Sky Conditions and Flat Terrain: For a "very clear" sky condition, when $\tau_{\mathrm{uw}}(\lambda,t) \approx 1$, $\tau_{\mathrm{dw}}(\lambda,t) \approx 1$ and $L_{\mathrm{airlight}}(\lambda,t) = E_{dif}(\lambda,t) \approx 0$[5], [93], if term $L_{adj}(p,\lambda,t)$ is also ignored, i.e., $L_{adj}(p,\lambda,t) \approx 0$, while, due to a flat terrain hypothesis, $E_{ter}(p,\lambda,t) \approx 0$ and incident angle $\varphi(p,t)$ equals the sun zenith angle $\theta_z(p,t)$, i.e., $\varphi(p,t) \equiv \theta_z(p,t)$, then the $SURF(p,\lambda,t)$ coefficient computed via (4) is approximated as follows [5], [110]:

$$TOARF(p,Band,t) = \pi \cdot d_{SE}(t)^2 \cdot \frac{TOARD(p,Band,t)}{ESUN(Band) \cdot \cos(\vartheta_z(p,t))} \in [0,1] \quad (5)$$

where the sensor-specific band-averaged solar spectral irradiance, $ESUN(Band)$, can be found in the sensor characteristics specification or in the image metadata file, while the acquisition time t and the solar zenith angle θ_z are provided with the image metadata file (with extension.IMD in the WV-2 image case).

To recapitulate, $TOARF =$ (5) provides an approximation of $SURF =$ (4) when atmospheric effects are ignored together with topographic effects (flat terrain hypothesis), whereas BRDF effects are omitted because the Lambertian surface hypothesis holds.

In [5], it is shown that when $SURF =$ (4) is approximated with $TOARF =$ (5), the latter is affected by two error terms due to atmospheric effects that tend to compensate for each other. Across wavelengths, this property improves the effectiveness of TOARF as an estimator of the true SURF values.

In addition, when wavelength λ increases, then TOARF provides a better approximation of SURF [5]. It is well known that light scattering due to atmospheric conditions (haze, consisting of gas molecules and water droplets) and aerosols (consisting of liquid droplets and solid particles suspended in the atmosphere and generated by either natural or anthropogenic sources) is inversely proportional to the energy wavelength λ, i.e., shorter wavelengths of the spectrum are scattered more than the longer

$$[0,1] \ni SURF(p,\lambda,t) = \pi \cdot d_{SE}(t)^2 \cdot \frac{1}{\tau_{uw}(\lambda)} \cdot \frac{(TOARD(p,\lambda,t) - L_{airlight}(\lambda,t) - L_{adj}(p,\lambda,t))}{\tau_{dw}(\lambda,t) \cdot ESUN(\lambda) \cdot \cos(\varphi(p,t)) + E_{dif}(\lambda,t) \cdot s + E_{ter}(p,\lambda,t)} \quad (4)$$

wavelengths. Thus, a visible blue (B) channel is affected by scattering across all atmospheric conditions ranging from "very clear" (where scattering is proportional to a factor λ^{-4}) to "very hazy" (where scattering is proportional to a factor $\lambda^{-0.5}$) and cloudy (where complete scattering occurs, proportional to a factor λ^{0}) [93]. On the contrary, in the medium infrared (MIR) wavelengths the amount of atmospheric scattering is known to be "quite small except for "very hazy" atmospheres and can be considered negligible" [93, p. 476].

To summarize, atmospheric effects can be omitted (ignored), i.e., $SURF = (4) \approx TOARF = (5)$ when: 1) visible wavelengths are acquired in "very clear" or "clear" sky conditions and 2) the NIR and MIR portions of the electromagnetic spectrum are acquired in all various atmospheric conditions unless sky is "very hazy."

Typically, a dark object subtraction technique is recommended to reduce atmospheric effects due to the upwelling path radiance [5], [93], [110]. In practice, when (4) is adopted assuming that $\tau_{\mathrm{uw}}(\lambda, t) \approx 1$ and $\tau_{\mathrm{dw}}(\lambda, t) \approx 1$, while terms $E_{dif}(\lambda, t)$, $E_{ter}(p, \lambda, t)$, and $L_{adj}(p, \lambda, t)$ are ignored and the flat terrain condition holds such that the incident angle $\varphi(p, t) \equiv \theta_z(p, t)$, then (5) can be replaced by the following approximation of (4), namely:

$$TOARF(p, Band, t) = \pi \cdot d_{SE}(t)^2 \cdot \frac{(TOARD(p, Band, t) - L_{airlight}(Band, t))}{ESUN(Band) \cdot \cos(\vartheta_z(p, t))}. \quad (6)$$

If (6) is adopted in place of (5) then, by definition, $TOARF = (6) = 0$ for a dark object (blackbody), then the unknown variable $L_{airlight}(\lambda)$ is equal to the $TOARD = (1)$ value measured upon the blackbody [110].

4) Anisotropic BRDF Effect Correction from SURF Values into Spectral Albedo: The task of BRDF correction is to derive, for non-Lambertian surfaces, *spectral albedo (bi-hemispherical reflectance, BHR)* values, defined over all directions [104], from $SURF = (4)$ values, $TOARF = (5)$ or $TOARF = (6)$ values where the Lambertian surface assumption holds. For operational use, a LC-dependent BRDF anisotropy factor, $K_{anstrpc}(p, \lambda, t)$, needs to be calculated for each surface type, which accounts for the relation between measured $SURF(p, \lambda, t)$ values [86] or TOARF(p, λ, t) values [106] and the spectral albedo, such that

$$BHR\left(p, \lambda, t) = \frac{SURF(p, \lambda, t)}{K1_{anstrpc}(p, \lambda, t, \text{surface type}}\right) \text{OR} \frac{TOARF(p, \lambda, t)}{K2_{anstrpc}(p, \lambda, t, \text{surface type})} \quad (7)$$

where the BRDF factor $K_{anstrpc}$ $(p, \lambda, t,$ surface type) is estimated from an appropriate either statistical or physical surface type-specific BRDF model [83]–[86].

A concern on validity may arise from the fact that a surface BRDF model should not be confused with a TOA (at-sensor) BRDF model. However, "some studies have shown that the anisotropic properties of the surface dominate over the impact of the atmospheric anisotropy for sensor spatial resolution finer than tens of kilometers. Apart from satellite resolution the relative importance of the surface and the atmosphere in effecting TOA BRDFs depends on climate regime. It is expected that atmospheric effects on TOA BRDFs are larger for a humid and hazy atmosphere than for a dry and clean one" [104].

A short survey of existing literature on BRDF effect correction in RS images is proposed below.

In [112], BRDF factor effects on a dried lake in Tuz Golu, Turkey (refer to this section above), were assessed by the National Physical Laboratory (UK) using the Gonio-Radiometric Spectrometer System. BRDF factor measurements were performed in the timeframe of the satellite overpasses to be used as input to a radioactive transfer code in the vicarious calibration process, where a full sequence of measurements at viewing angles: 10°, 20°, 30° takes 10 min. The spectrometer used with the goniometer system operated over the spectral range 400–1300 nm. As stated in this paper: "The results show that there are dramatic changes in BRDF effects for small changes in solar zenith angle of approximately 0.25 reflectance factor (i.e., there is a difference in surface reflectance of 0.25 when reflectance ranges in [0, 1]) at 650 nm over a 30 degree change in measurement angle. This could be the result of the surface structure of the salt flat."

In the ATCOR software product (refer to Table I), the physical model-based radiometric processing chain, shown in Fig. 1, requires as input three major entities [83]–[86].

- Atmospheric look-up tables created using a radiative transfer code (e.g., MODTRAN) and some initial knowledge about the state of the atmosphere (e.g., aerosol model).
- DSM and/or DTM data and their derived quantities such as terrain slope, aspect, and skyview factor. Required DSM constraints are mentioned above in Section IV-A2.
- TOARD calibrated and geocoded image data, stored in raw geometry including all geometric information (e.g., pixel location, solar and sensor geometry, etc.).

The radiative transfer parameters required for the inversion of the at-sensor radiance $TOARD = (1)$ to calculate the atmospheric parameters and the spectral albedo, $BHR = (7)$, typically comprise 5 up to 7 ancillary input data dimensions [83]–[86], [104], [113].

- Sensor view and azimuth angles, solar zenith, and azimuth angles.
- AOT (depth). In the ATCOR commercial software product it is computed with a standard automatic AOT retrieval approach which is class specific and works in areas with dark surfaces, i.e., it requires preliminary classification, performed by the ATCOR-SPECL pre-classifier, which is the only existing symbolic syntactic alternative to the SIAM™ software product found in commercial software toolboxes [83]–[85], [115] (refer to Section II-G). In the context of the ATCOR software toolbox, the ATCOR-SPECL sub-system is considered a by-product. An atmosphere visibility (in kilometers or, *vice versa*, a dimensionless optical depth) can be selected, e.g., a summer atmosphere visibility of approximately 30 km.
- Aerosol model, e.g., maritime/rural/urban/desert aerosol model.
- Atmospheric water vapor amount, which is class specific and requires preliminary classification, performed by the ATCOR-SPECL pre-classifier [85], [115]. For example,

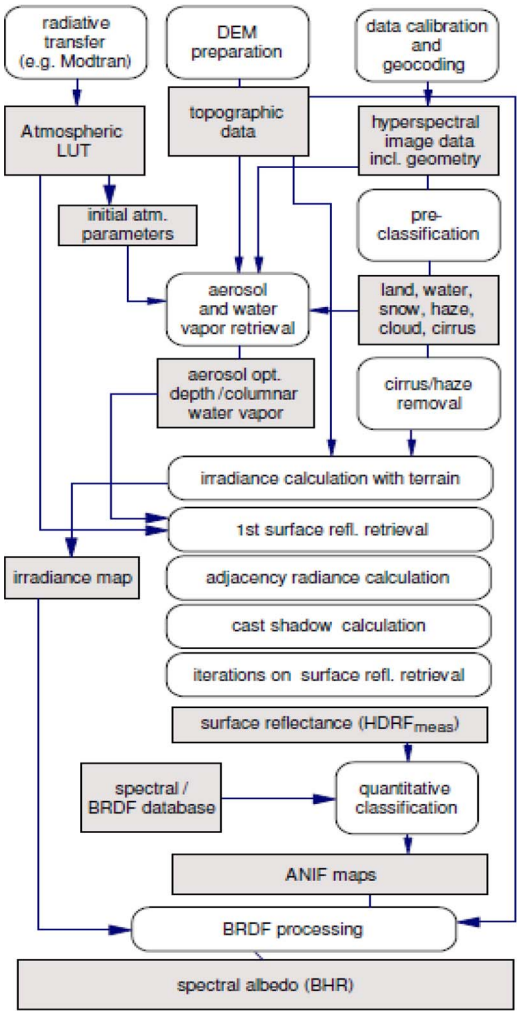


Fig. 1. Same as in [86], courtesy of Daniel Schläpfer, ReSe Applications Schläpfer. Complete atmospheric correction and radiometric normalization scheme implemented in an "augmented" version of the ATCOR software product [83]–[86], capable of transforming sensory data into surface reflectance values and, next, spectral albedo. Processing blocks are represented as circles, and output products are represented as rectangles. In this physical model-based workflow, categorical variables, generated as output by multiple preliminary classification stages (e.g., refer to blocks identified as "pre-classification" and "quantitative classification"), are required for continuous bio-physical variables estimation to be conducted on a stratified categorical variable-specific basis. The so-called pre-classification stage is implemented as the non-adaptive, rule-based, spectral classifier SPECL, considered as a by-product in the ATCOR software toolbox. Among existing commercial software products, such as those listed in Table I, the ATCOR-SPECL pre-classifier appears as the only alternative to the symbolic, syntactic, static SIAM™ [5]–[17]. In practice, SIAM™ is eligible for replacing the ATCOR-SPECL sub-system in the spectral albedo estimation workflow shown in this figure.

a water vapor column in centimeters can be selected from a known set of instances, e.g., tropical conditions/ midlatitude summer/dry summer, spring, or fall/dry desert or winter.

- DSM or DTM, to derive quantities such as terrain slope, aspect and skyview factor.
- Wavelength.

In the ATCOR software product [83]–[86], where the standard automatic method of AOT retrieval works in areas with dark surfaces and is extended to the automatic selection of standard aerosol models, there are two main sources of errors in surface albedo estimation, particularly for VHR imaging sensors: first, the spectral correlation of MIR (also called short-wave IR, SWIR), visible, and NIR channels is uncertain over vegetation and water surface types and, second, dark objects highly depend on the surface cover type. Therefore, the aerosol characterization is currently based on standard values in non-vegetated areas, although new methods are still under investigation over non-vegetated surfaces, e.g., bright surfaces.

Two considerations about the estimation of $BHR = (7)$ stem from this short analysis of the physical model-based ATCOR radiometric processing workflow shown in Fig. 1 [83]–[86]: *1)* (7) *is LC class-specific* [114] *and 2) in common practice, the solution of* (7), i.e., *the estimation of a continuous physical variable from sensory data, requires as input, at several stages of the data processing workflow, categorical variables belonging to preliminary classification maps automatically generated from the same RS image to be radiometrically corrected (see Fig. 1). In other words, equivalent to two sides of the same coin, categorical variables* (e.g., *LC and LCC maps*) *and continuous variables* (e.g., *spectral albedo, LAI, green biomass*, etc.) *should be estimated from RS images alternately and iteratively. It means that, in a RS image processing workflow conceived as a hybrid inference feedback system, like the stratified TOC algorithm proposed in* [10] (*refer to Section IV-A2*), *continuous variables are estimated on a categorical (stratified) basis while (enhanced) categorical variables are estimated from (enhanced) continuous variables in addition to prior knowledge (refer to Section II-B).*

In [104], physical model-based LC class-specific BRDF linear models are generated for four LC types observed by the AVHRR sensor as a function of four free parameters, namely, three geometric parameters, the solar zenith angle, satellite viewing zenith angle, and relative azimuth angle (defined as the difference between the Sun and the sensor azimuth angles), plus one physiological canopy parameter, the normalized difference vegetation index (NDVI) considered as a proxy of the LAI and the green biomass. Three linear model coefficients are empirically estimated as functions of NDVI for each LC type by minimizing the differences between the observed and the modeled reflectances using an optimization algorithm. Observations are collected in multiple RS images providing a proper sample of the geometric parameter 3-D space. The four target LCs are: barren, grassland, forest, and cropland. In these experiments, the satellite viewing zenith angle dependence of reflectance for barren and grassland is significantly stronger than for cropland and forest.

In [105], physically based BRDF models employing three free parameters to correct SURF values of five canopies (indigenous forest, exotic forest, scrub, pasture, and tussock) in AVHRR images are fitted statistically. To perform BRDF correction, these parameters can be used in conjunction with a vegetation map specifying proportions of these groups at any given location, and the correction applied as a linear combination of BRDF models.

In [106], physical model-based LC class-specific BRDF linear model parameters for radiometric normalization (registration) of multi-temporal Landsat images are estimated in image overlapping areas. LC class-specific BRDF linear model

TABLE VI
WV-2 AND QB-2 IMAGE SET, PROVIDED BY DIGITALGLOBE FOR TESTING PURPOSES IN THE FRAMEWORK OF THE 2011 DIGITALGLOBE EIGHT-BAND CHALLENGE, AND LANDSAT-7 ETM+ IMAGES SELECTED AS "MASTER" IMAGES FOR RADIOMETRIC REGISTRATION (RE-CALIBRATION) OF THE TWO "SLAVE" WV-2 IMAGES

Spaceborne test images of Brazilia (Brazil)		Acquisition date and time	Sensor position (azimuth, elevation in degrees)	Off-Nadir View Angle	Sun position (azimuth, elevation in degrees)	Spatial resolution (m)	Spectral resolution (μm) per band
'Slave' WorldView-2 (WV-2)	WV-2, T1 (green season)	2010-02-04, 13:26	SatAz = 104.6, SatEl = 70.8	17.0	SunAz = 95.0. SunEl = 61.3	1.84 at nadir, 2.4 at 20° off-nadir	1-Coastal Blue (Violet, CB): 0.400-0.450, 2-B: 0.450-0.510, 3-G: 0.510-0.580, 4-Yellow (Y): 0.585-0.625, 5-R: 0.630-0.690, 6-Red Edge (RE): 0.705-0.745, 7-NIR1: 0.770-0.895, 8-NIR2: 0.860-1.040
	WV-2, T2 (dry season)	2010-08-04, 13:26	SatAz = 104.3. SatEl = 70.9	17.0	SunAz = 39.4. SunEl = 48.1	Same as above	Same as above
QuickBird-2 (QB-2), T1 + 45 days (green season)		2010-03-16, 13:26	SatAz = 81.5. SatEl = 80.1	9.2	SunAz = 66.2. SunEl = 58.6	2.44	1-B: 0.445-0.510 , 2-G: 0.500-0.595, 3-R: 0.620-0.690, 4-NIR: 0.755-0.875
'Master' Landsat-7 Enhanced Thematic Mapper (ETM)+	ETM+, T1 (green season).	2010-02-03, 13:06.			SunAz = 95.91. SunEl = 56.59.	Bands 1-5, 7: 30, Band 6: 60.	1-B: 0.45-0.52, 2-G: 0.52-0.60, 3-R: 0.63-0.69, 4-NIR: 0.76-0.90, 5-MIR1: 1.55-1.75, 6(1)-TIR: 10.4-12.5, 6(2)-TIR: 10.4-12.5, 7-MIR2: 2.08-2.35
	ETM+, T2 (dry season).	2010-07-13, 13:06.			SunAz = 41.24. SunEl = 40.56.	Same as above	Same as above

parameters are three geometric variables, the solar zenith angle, the satellite viewing zenith angle, and the relative azimuth angle (refer to this section above), while the linear model coefficients to be estimated are two. First, $TOARF = (5)$ is computed (which assumes the surface be Lambertian), and, next, areas of cloud, smoke, recent fire scars, and crops (where considerable seasonal change are observed) are masked out (rejected) to minimize the real change component in overlap areas. Finally, in the image overlapping areas, masked to include the target LCs only, two surface type-specific BRDF linear model coefficients are estimated when the model geometric parameter 3-D space is properly sampled by multiple RS images acquired in a small range of time (to guarantee that target LC types do not change their pictorial properties during parameter estimation). The target non-Lambertian LC classes (strata) are: water, bare soil, woody, and non-woody vegetation.

In [114], the same empirical BRDF model employed in [106] is easily adapted for VHR imagery, where shadow (casted by trees and buildings, accounting for one quarter of the entire image area) detection and removal remains a very challenging problem.

5) Implemented Pre-Processing of the WV-2 and QB-2 Test Images: Provided by DigitalGlobe for testing purposes, the three VHR images acquired over the capital site of Brazilia (Brazil) in 2010 feature the acquisition parameters shown in Table VI. They encompass a WV-2 image acquired at T1, corresponding to the green season, a WV-2 image acquired at T2, corresponding to the dry season, and a QB-2 image acquired at $T1 + 45$ days, corresponding to the green season.

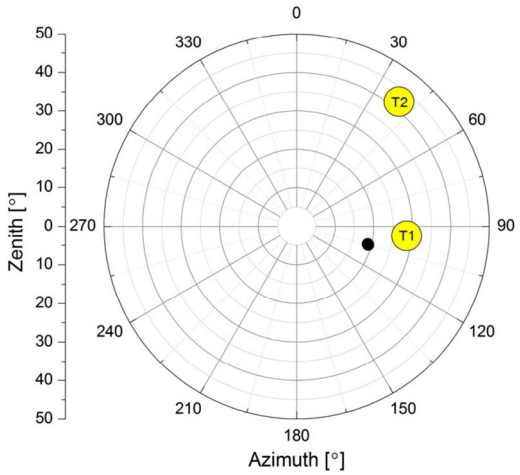


Fig. 2. Ground observed azimuth and zenith angles of the satellite positions (black circles, coincident as one) and sun positions (yellow circles) for the two test images WV-2 T1 and T2.

In the rest of this paper, the eight WV-2 bands are identified as follows: Band 1, 400–450 nm, Coastal Blue (Violet, CB); Band 2, 0.450–0.510 nm, visible Blue (B); Band 3, 0.510–0.580 nm, visible Green (G); Band 4, 0.585–0.625 nm, visible Yellow (Y); Band 5, 0.630–0.690 nm, visible Red (R); Band 6, 705–745 nm, Red Edge (RE); Band 7, 0.770–0.895 nm, NIR1; Band 8, 0.860–1.040 nm, NIR2.

The four QB-2 bands are identified as follows: Band 1, 0.450–0.520, visible Blue (B); Band 2, 0.520–0.600, visible

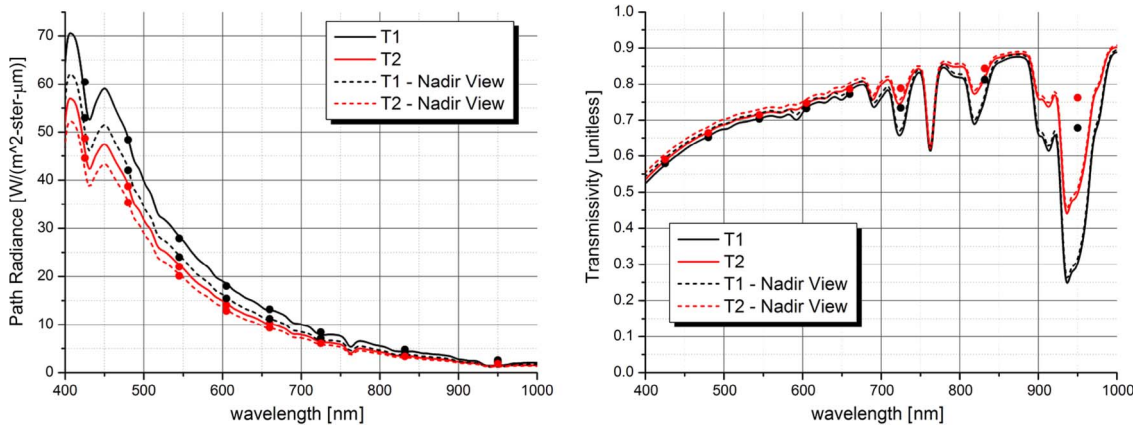


Fig. 3. (a) Upwelling path radiance for images WV-2 T1 and T2 in comparison with their simulated nadiral observations. (b) Downwelling transmissivity, $\tau_{dw} \in [0, 1]$, for images WV-2 T1 and T2 in comparison with their assumed nadiral observations.

Green (G); Band 3, 0.630–0.690, visible Red (R); Band 4, 0.760–0.900, NIR.

Table VI shows that the WV-2 data set is particularly suitable for multi-temporal analysis as the viewing geometries are nearly identical between the two acquisitions, although the Sun position changes significantly from East to North-East. Due to the difference in Sun positions, then differences in atmospheric effects (e.g., upwelling path radiance component of the at-sensor radiance), together with surface type-specific BRDF effects, are expected to affect the two test WV-2 images.

Fig. 2 illustrates the ground observed azimuth and zenith angles of the satellite positions and sun positions for the two test images WV-2 T1 and T2.

Fig. 3(a) illustrates the continuous curve values of $L_{airlight}(\lambda, t)$ calculated by MODTRAN simulations for the WV-2 T1 and T2 images, together with the two dotted curves of the two WV-2 T1 and T2 images simulated with a sensor nadir view (i.e., satellite viewing zenith angle equals zero [104]), e.g., to mimick a Landsat sensor viewing geometry. In the two WV-2, T2 and T1 continuous curves the atmospheric visibility is estimated equal to 31 km and 34 km, respectively, where a 30 km atmospheric visibility is a standard approximation in most applications for clear-sky conditions. Between the T1 and T2 continuous curves, differences in airlight are up to 25%, the former showing larger upwelling path radiance values. Differences between continuous (off-nadir view) and dotted (nadir view) curve pairs across wavelengths are up to 15% for T1 and 10% for T2, where larger airlight values are those acquired in the off-nadir view. It means that in comparison with Landsat images, WV-2 T1 and T2 values are larger by, respectively, a 15% and a 10% factor. Fig. 2(b) illustrates the downwelling atmospheric transmissivity for the same four cases mentioned above, where the biggest differences between the two test images WV-2 T1 and T2 are found in the the RE and NIR2 bands. These differences are due to the different concentration of estimated water vapor between the wet (T1) and dry (T2) seasons.

a) Atmospheric correction: Atmospheric correction typically requires as input ancillary data (summary statistics, e.g., AOT, water vapor amount, aerosol model, etc. [83]–[86], refer to Section IV-A2), to be collected at several locations within the RS image footprint at the time of acquisition of the RS image, but are rarely available in practice. Hence, the problem of atmospheric correction is inherently ill- or poorly posed, i.e., it is difficult to solve and require user's supervision to make it better posed [5]. For example, in [105], to improve the atmospheric correction performed by the 6S atmospheric transmission model [107], monthly climatological mean profiles of pressure, temperature, water vapor, and ozone were collected by ozone probes.

The present authors have repeatedly observed that EO images radiometrically calibrated into SURF values and delivered by EU institutions, data providers, RS scientists, and practitioners are often affected by spectral distorsion. This means that SURF spectra extracted from these atmospherically corrected images lack physical meaning, i.e., these estimated spectra disagree with reference SURF spectra, measured at the ground level, found in the existing literature (e.g., refer to [94, p. 273] or in public domain spectral libraries [83]–[86] (refer to Section II-G). This lack of physical meaning is particularly true for estimated SURF range extrema, e.g., pixel values belonging to either very bright (e.g., snow, light-tone buildings, etc.) or very dark (e.g., water) image-objects.

Based on these considerations, no empirical atmospheric correction is adopted to estimate $SURF = (4)$ values, but $TOARF = (5)$ values are computed instead. This conservative choice is justified by the fact that the prior knowledge base of the syntactic SIAM™ preliminary classifier consists of a reference dictionary of spectral signatures in TOARF values, where TOARF $\supseteq$ SURF, such that TOARF $\approx$ SURF + atmospheric noise (refer to Section II-G). This means that SIAM™ is knowledgeable to cope with noisy RS data, namely, it is capable of recognizing surface types through atmospheric noise, like haze and thin clouds [5]–[17] (refer to Section II-G).

b) TOC: In general, in case of VHR imagery, the TOC requirements specification mentioned in Section IV-A2 becomes almost impossible to fulfill. In the words of ATCOR [86]: "at spatial resolutions down to 0.5 m, the slope of surfaces can no longer be easily defined and no generic irradiance correction can be applied, e.g. in forests or settlements. New models for radiometric surface representation would be required but no generic solution is currently available."

In this paper, no DSM of the target surface area could be found to fulfill the aforementioned TOC constraints. Hence, no TOC approach is applied.

c) BRDF effect correction: No required library of sensor-specific multiple views of a target surface type is available for physically based BRDF model fitting. Hence, no BRDF model can be applied.

This means that, to make the two test WV-2 images comparable (e.g., for LCC detection) by reducing the inherent spread (variance) of reflectances acquired through time in different acquisition conditions upon the same non-Lambertian surface type, a statistical model-based approach must be adopted for empirical radiometric registration (matching) of the two test WV-2 images to a reference data set, so that it appears as if they were acquired under the same set of acquisition conditions (refer to the introduction to Section IV-A).

d) Assessment of the information content of the two test WV-2 images radiometrically calibrated into TOARF values: Table VI shows that, in comparison with the QB-2 spectral channels, the WV-2 band 1 (CB), band 4 (Y), band 6 (RE), and band 8 (NIR2) can be considered "*new*," while "*traditional*" WV-2 channels, (approximately) in common with the QB-2 spectral resolution, are: band 2 (B), band 3 (G), band 5 (R), and band 7 (NIR1).

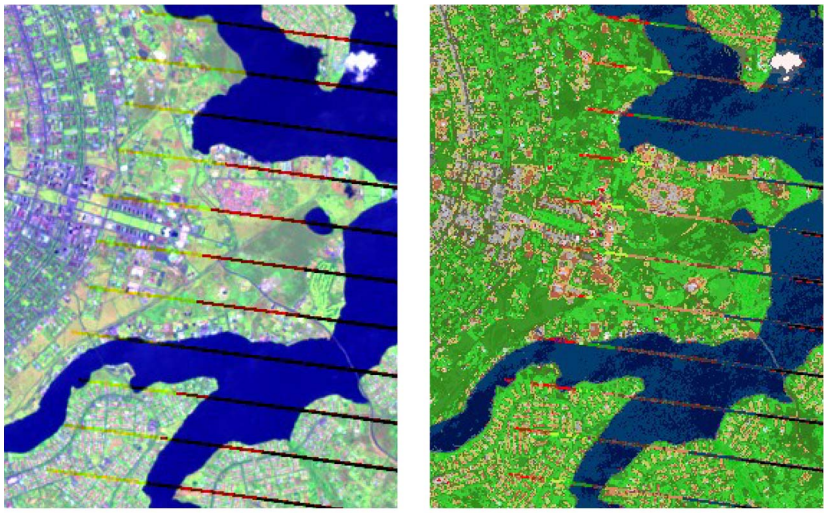
In the two WV-2 test images radiometrically calibrated into $TOARF = (5)$ values, it is observed that, across their eight bands, the average inter-band correlation is ≥ 0.933 for image T1 and ≥ 0.939 for image T2. The traditional threshold adopted by Congalton and Green to consider correlation high is 0.8 [51]. Hence, in the two test WV-2 images of Brazilia acquired in the wet (T1) and dry (T2) seasons of the year 2010, the average inter-band correlation is high, i.e., their information redundancy is statistically high. The same consideration holds true for the WV-2 image of the downtown area of Rome, Italy, acquired on 2009-12-10 and downloaded as a product sample from the DigitalGlobe website [116], whose average inter-band correlation is > 0.9. Based on these considerations, the conjecture made here is that the average inter-band correlation of the eight-band WV-2 channels is expected to be high (> 0.8 [51]) across time and space. In other words, the eight-band WV-2 inter-band information redundancy is expected to be high irrespective of time and space, although it is important to consider that correlation is insensitive (invariant) under a linear transformation of the two random variables at hand (e.g., reflectances generated from the same target surface type in two different spectral bands).

In the existing RS literature, the aforementioned "*new*" WV-2 channels have been considered relevant based on evidence collected from statistical classification approaches [117], [118]. Additional physical model-based considerations about the potential discrimination capability of the "*new*" WV-2 channels CB and RE can be inferred from the RS literature as summarized below.

- About the discrimination capabilities of the WV-2 band 6 (RE) in 705–745 nm (see Table VI). In [119], shifts in the Red Edge Position (REP) are found to improve separation: 1) between broadleaved species from coniferous and grassland and 2) between coniferous of different ages. Unfortunately, shifts in REP can be detected by hyperspectral sensors exclusively. In [120], two Red Edge (RedE1 and RedE2) plus one visible Red band of a hyperspectral MIVIS sensor are selected for tree species discrimination: Red band ranges in 650–670 nm, RedE1 in 690–710 nm and RedE2 in 730–750 nm. It is noteworthy that the WV-2 RE band (refer to Table VI) overlaps with both the two aforementioned MIVIS bands RedE1 and RedE2. For vegetation detection in general, it is common knowledge to consider the Landsat MIR1 channel, which is sensitive to both vegetation moisture content and soil moisture, as the best Landsat band overall, superior to bands NIR, MIR2 and visible (see Table VI) [143]. In [121], where reference samples of vegetated LCs do not include crops stressed by limited soil water availability, most valuable bands for the estimation of LAI are considered the NIR band, followed by the RE, MIR, and visible bands. In these experiments, the condition that water required for crop irrigation is fully available diminishes the discrimination capability of the MIR band upon the reference vegetation sample at hand, while the NIR band remains superior to the RE channel.
- About the discrimination capabilities of the WV-2 band 1 (CB) in 400–450 nm (see Table VI). It comprises Moderate Resolution Imaging Spectroradiometer (MODIS) bands Violet 1 (V1, 0.405–0.420 μm) and V2 (0.438–0.448 μm) which are considered relevant by oceanographers together with MODIS bands Blue 1 (B1, 0.459–0.479 μm) and B2 (0.483–0.493 μm). In comparison with the WV-2 band 2 (B) channel ranging in 450–510 nm (see Table VI), the WV-2 band 1 CB should see further into the water and support bathymetric studies around the globe. To hold true, this potential improvement requires: 1) clear water and, simultaneously, 2) very clear atmospheric conditions to reduce atmospheric scattering [93].

To recapitulate, the "*new*" WV-2 channels, CB, Y, RE, and NIR2 (see Table VI), are expected to be useful, i.e., statistically uncorrelated to the "*traditional*" WV-2 bands R, G, B, and NIR1, at a local image scale and/or in EO circumstances whose relevance can be considered high by application domain experts (e.g., oceanographers), but whose statistical occurrence can be considered low, or unlikely (e.g., the RE band may be useful in the separation of specific tree species found in very localized forest stands [119], [120]; the CB band requires clear coastal water to exploit its penetration capability and, simultaneously, very clear atmospheric conditions to reduce its atmospheric scattering, refer to this section above).

For example, in the two test, WV-2 images of Brazilia available in this work, at a qualitative level of visual analysis none of the "new" WV-2 bands appears capable of providing additional LC class discrimination capabilities. For example, water appears turbid and cannot be penetrated by the WV-2 band 1 CB significantly better than the WV-2 band 2 B does. In addition, no significant wood/forest tree population does exist in the two WV-2 test images to be discriminated by the WV-2 band 6 RE any better than the WV-2 band 7 NIR1 does.



(a) (b)

Fig. 4. (a) Subset of interest of a Landsat-7 ETM+ image at time T1, 30 m spatial resolution, acquisition date 2010-02-03, radiometrically calibrated into TOARF values, depicted in false colors (R: band 5, G: band 4, B: band 1). Default image histogram stretching: ENVI linear stretching 2% [138]. (b) L-SIAM™ preliminary map of the Landsat-7 ETM+ image at time T1 shown in Fig. 4(a). Spectral categories are depicted in pseudo colors. Map legend: refer to Table III.

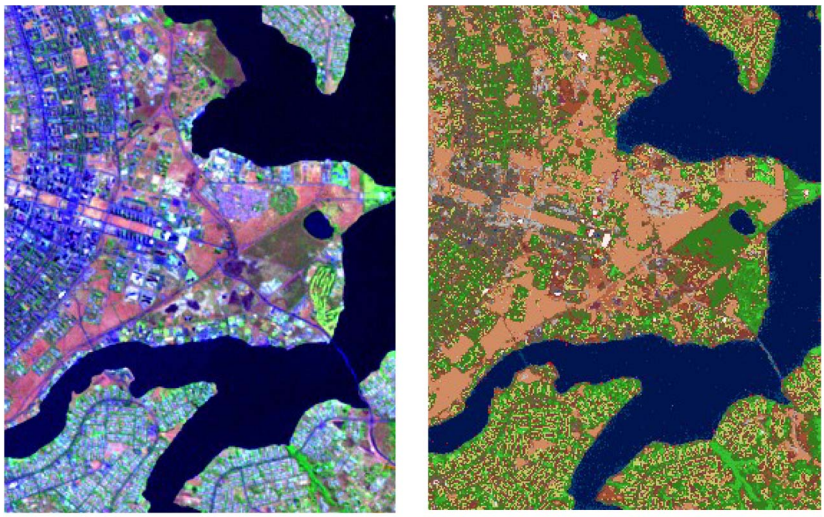
Overall, the aforementioned WV-2 data-specific observations can be considered quite obvious. They are perfectly in line with the RS common practice. For example, it is well known that inter-band correlations of visible channels Red, Green, and Blue acquired by existing spaceborne optical sensors (e.g., Landsat, MODIS, etc., see Table VI), irrespective of their spatial resolution, are typically high (> 0.8). As another example, in hyperspectral images, the average inter-band correlation is typically high. For example, in an airborne, 285 band (ranging from visible Blue to MIR), 1.8 m resolution Airborne Prism Experiment hyperspectral image of Baden, Switzerland, acquired on 2011-06-26 (courtesy of Daniel Schläpfer, ReSe Applications Schläpfer), the average inter-band correlation is > 0.8. It is superfluous to point out that the high inter-band correlation of hyperspectral images does not mean at all that hyperspectral images cannot be very useful. It only means that, in the RS common practice, RS-IUSs adopt a hyperspectral image-specific feature selection first stage suitable for spectral dimensionality reduction, e.g., principal component analysis, spectral end-member detection, etc. [85], [86], [119].

Based on the aforementioned WV-2 image-specific qualitative and quantitative observations, supported by theoretical considerations of general validity selected from the existing literature, *this experimental session selects the existing four-band Q-SIAM™ software product implementation* (refer to Table II) *to map the two test WV-2 images based on "traditional" bands B, G, R, and NIR1 exclusively*, i.e., *the WV-2 "new" bands, CB, Y, RE, and NIR2, are ignored in these experiments* (because their contribution in terms of discrimination capability is considered negligible in the two test images at hand). Of course, this experimental strategy is adopted with the sole objective of testing a map quality assessment protocol. It does not mean at all that, in general, the "new" WV-2 bands CB, Y, RE, and NIR2 are considered worthless. For example, it would be perfectly reasonable to implement an *ad-hoc* second-stage WV-2 sensor-specific spectral rule set to exploit the potential discrimination capabilities of the "*new*" WV-2 channels CB, Y, RE, and NIR2, in series with the standard Q-SIAM™ preliminary classification first stage exploiting as input the "traditional" WV-2 bands R, G, B, and NIR1 which are shared with QB-2 and a variety of other sensors (refer to Table V).

e) Radiometric registration of "slave" off-nadir 2-m resolution WV-2 images to "master" nadir-view 30-m resolution Landsat images: To reduce the inherent spread (variance) of the WV-2 data acquired through time in different acquisition conditions upon the same non-Lambertian surface type, including permanent reflectors (such as asphalts and concrete surface areas), a statistical model-based approach is adopted for empirical radiometric registration (matching) of the test off-nadir 2-m resolution WV-2 T1 and T2 images to a reference data set consisting of two nadir-view 30-m resolution Landsat images acquired (approximately) at time T1 and T2, respectively.

Despite the recent malfunction in the Landsat-7 ETM+ sensor and the recent termination of the long-lived Landsat-5 TM mission, Landsat data continue to have tremendous scientific utility [2]. It is well known that Landsat sensors are well-behaved and stable. For example, the Landsat-5 TM and Landsat-7 ETM+ radiometric calibration uncertainties of the at-sensor spectral radiances are both around 5%. ETM+ is the most stable of the Landsat sensors, changing by no more than 0.5% per year in its radiometric calibration [122]. In addition, since the Landsat satellites are not adjustable nadir viewing, they tend to minimize BRDF effects. Thus, two cloud-free Landsat-7 ETM+ images were selected from the USGS Global Visualization Viewer (http://glovis.usgs.gov/) as a reference for the test WV-2 image pair for inter-sensor data calibration purposes. Acquisition parameters of the two reference ("master") Landsat-7 ETM+ images are described in Table VI.

The two "master" 30-m resolution Landsat-7 ETM+ images, radiometrically calibrated into $TOARF = (5)$ values, are shown in Figs. 4(a) and 5(a), together with their L-SIAM™ maps shown in Figs. 4(b) and 5(b), respectively. Next, the two "slave" 2-m resolution WV-2 T1 and WV-2 T2 images,



(a) (b)

Fig. 5. (a) Subset of interest of a Landsat-7 ETM+ image at time T2, 30 m spatial resolution, acquisition date 2010-07-13, radiometrically calibrated into TOARF values, depicted in false colors (R: band 5, G: band 4, B: band 1). Default image histogram stretching: ENVI linear stretching 2% [138]. (b) L-SIAM™ preliminary map of the Landsat-7 ETM+ image at time T2 shown in Fig. 5(a). Spectral categories are depicted in pseudo colors. Map legend: refer to Table III.

radiometrically calibrated into $TOARF = (5)$ values, are re-calibrated with respect to their "master" 30-m resolution Landsat ETM+ images as described below.

Four regions of interest (ROIs) are identified across spatial resolutions in the overlapping stack of two Landsat-7 ETM+ (master) and two WV-2 (slave) images. Based on a photointerpetation process in combination with thematic evidence provided by the L-SIAM™ maps of the two Landsat-7 ETM+ images [refer to Figs. 4(b) and 5(b)], the following surface composition through time is assigned to each ROI.

- ROI1 (constant water). T1 (namely, green season): Water, T2 (namely, dry season): Water.
- ROI2 (vegetation with within-class variance due to seasonality). T1: Vegetation (grassland), T2: Vegetation (shrub rangeland, whose LAI is inferior from that of grassland).
- ROI3 (permanent scatterer). T1: Flat roof in white building, T2: Flat roof in white building.
- ROI4 (inter-class transition). T1: Vegetation (shrub rangeland), T2: Barren land.

Stemming from a radiometric registration approach, radiometric correction linear parameters for the WV-2 band 2 B, band 3 G, band 5 R, and 7 NIR1 (refer to Section IV-A5d) at time T1 and T2 are shown in Table VII. These radiometric correction parameters reveal that the WV-2 T1 image needs a stronger linear correction with respect to its "master" ETM+ T1 image than the WV-2 T2 image with respect to its "master" ETM+ T2 image. This agrees with the physical model-based estimate of atmospheric effects proposed in Section IV-A5a.

The effectiveness of the accomplished radiometric registration of the two "slave" 2-m resolution WV-2 images to match the two "master" 30-m resolution Landsat images is proved by the standard deviation of the absolute differences in reflectance through time of permanent reflectors (e.g., ROI1 as permanent water, ROI3 as building roof, etc.), located in the overlapping areas of the two re-calibrated WV-2 images at time T1 and T2, which is significantly reduced, from 10% to 50%, with respect to the standard deviation of the absolute differences collected in the original WV-2 image pair radiometrically calibrated into $TOARF = (5)$ values. This is the first source of independent evidence confirming the consistency of the radiometric registration process applied to the two WV-2 images.

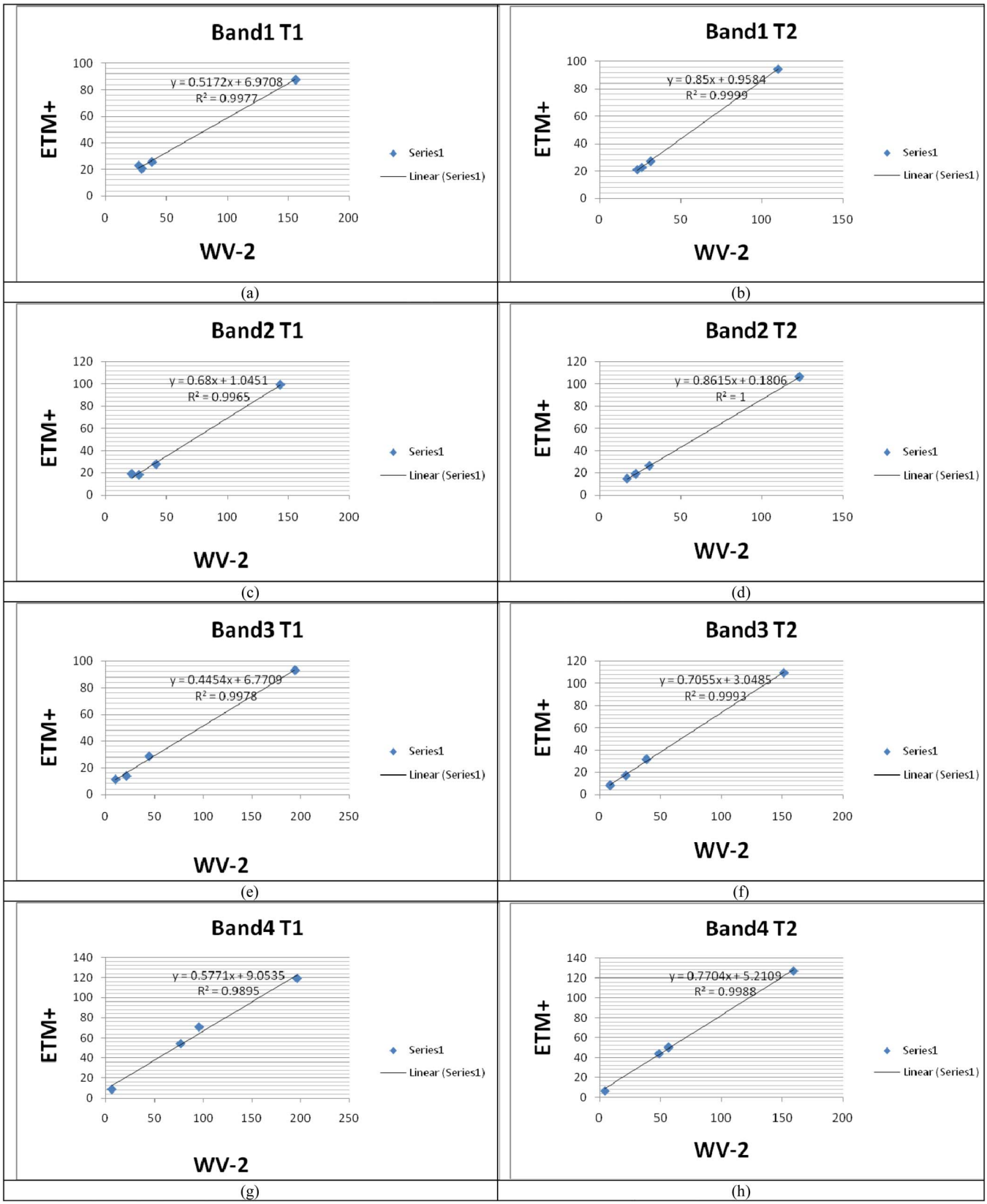
B. Q-SIAM™ Preliminary Classification of the WV-2 and QB-2 Test Images

The Q-SIAM™ output products generated from the re-calibrated WV-2 image at time T1, shown in Fig. 6(a), are depicted in Figs. 6(b) and 7. Those generated from the re-calibrated WV-2 image at time T2, shown in Fig. 8(a), are depicted in Figs. 8(b) and 9. In addition to preliminary classification maps at different semantic granularities, automatic Q-SIAM™ output products are described as follows:

- An eight-adjacency cross-aura map generated from the preliminary classification map (see Fig. 7(c) generated from Fig. 7(b) at time T1 and Fig. 9(c) generated from Fig. 9(b) at time T2, respectively). It highlights contours of symbolic image-objects automatically detected in the preliminary classification map domain (refer to footnote 1) [5]–[17].
- As an example of a categorical stratum, a binary vegetation mask is extracted from the preliminary classification map (see Fig. 7(d) generated from Fig. 7(b) at time T1 and Fig. 9(d) generated from Fig. 9(b) at time T2, respectively).

When compared with the "master" L-SIAM™ map pair shown in Figs. 4(b) and 5(b), the two Q-SIAM™ preliminary maps shown in Figs. 6(b) and 8(b) appear overall (qualitatively) consistent, such as overall consistent appear their inter-map local differences. For example, while moving from the green to the dry season, differences between the Q-SIAM™ preliminary maps of the two WV-2 images, equivalent to differences between Figs. 6(b) and 8(b), appear characterized by temporal

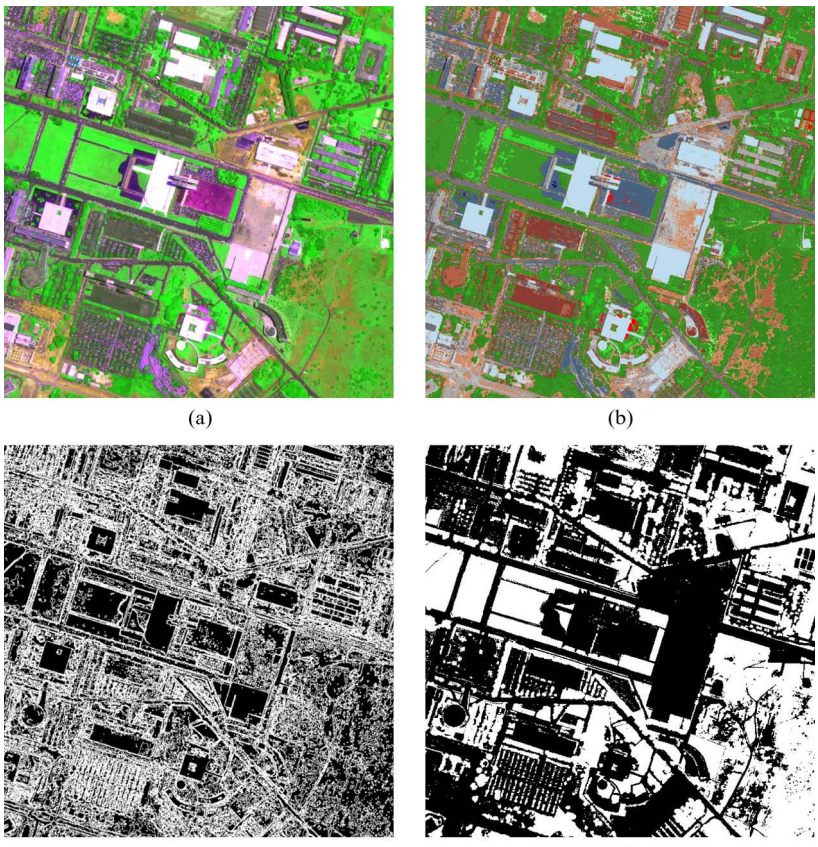
TABLE VII
RADIOMETRIC REGISTRATION APPROACH: BAND-SPECIFIC RADIOMETRIC CORRECTION LINEAR PARAMETERS ESTIMATED FOR THE FOUR BANDS B, G, R AND NIR1 OF TWO "SLAVE" WV-2 IMAGES, ACQUIRED AT TIME T1 AND T2 AND RADIOMETRICALLY CALIBRATED INTO TOARF VALUES, TO MATCH TWO "MASTER" LANDSAT-7 ETM+ IMAGES, ACQUIRED AT TIME T1 AND T2 AND RADIOMETRICALLY CALIBRATED INTO TOARF VALUES



shifts in vegetation labels equivalent to transitions from superior (e.g., "large") to inferior (e.g., "average" or "low") LAI values (refer to the Q-SIAM™ map legend shown in Table IV). This is the second source of independent evidence confirming the consistency of the radiometric registration process applied to the two WV-2 images.

(a) (b)

Fig. 6. (a) WV-2 T1 image, 2 m spatial resolution, acquisition date 2010-02-04, radiometrically calibrated into TOARF values and re-calibrated according to a "master" Landsat-7 ETM+ image shown in Fig. 4(a), depicted in false colors (R: 5, G: 7, B: 2). Default image histogram stretching: ENVI linear stretching 2% [138]. Compare this WV-2 image with its reference Landsat image shown in Fig. 4(a). (b) Q-SIAM™ preliminary map of the WV-2 T1 image shown in Fig. 6(a). Spectral categories are depicted in pseudo colors. Map legend: see Table IV. Compare this Q-SIAM™ map with the L-SIAM™ shown in Fig. 4(b), generated from the reference Landsat image depicted in Fig. 4(a).



(a) (b)

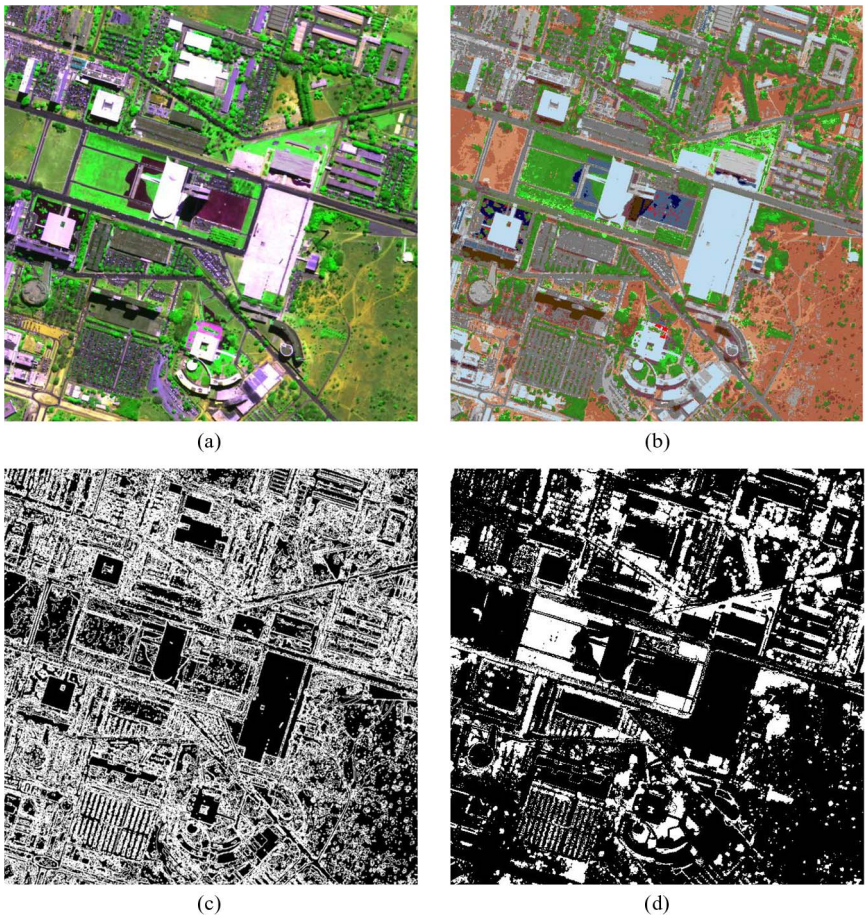
(c) (d)

Fig. 7. (a) Zoom of the WV-2 T1 image, 2 m spatial resolution, acquisition date 2010-02-04, radiometrically calibrated into TOARF values and shown in Fig. 6(a), depicted in false colors (R: 5, G: 7, B: 2). Default image histogram stretching: ENVI linear stretching 2% [138]. (b) Zoom of the Q-SIAM™ preliminary map, shown in Fig. 6(b), of the WV-2 T1 image shown in Fig. 7(a). Spectral categories are depicted in pseudo colors. Map legend: see Table IV. (c) 4-adjacency cross-aura measure generated from the Q-SIAM™ preliminary map, shown in Fig. 7(b), of the WV-2 T1 image shown in Fig. 7(a). Cross-aura values range in {0, 4}. (d) Binary vegetation mask generated from the Q-SIAM™ preliminary map, shown in Fig. 7(b), of the WV-2 T1 image shown in Fig. 7(a).

The test QB-2 image acquired at time $T1 + 45$ days, radiometrically calibrated into $TOARF = (5)$ values, is shown in Fig. 10(a). The Q-SIAM™ output products generated from the radiometrically calibrated QB-2 image at time T1 shown in Fig. 10(a) are depicted in Fig. 10(b)–(d).

(a) (b)

Fig. 8. (a) WV-2 T2 image, 2 m spatial resolution, acquisition date 2010-08-04, radiometrically calibrated into TOARF values and re-calibrated according to a "master" Landsat-7 ETM+ image shown in Fig. 5(a), depicted in false colors (R: 5, G: 7, B: 2). Default image histogram stretching: ENVI linear stretching 2% [138]. Compare this WV-2 image with its reference Landsat image shown in Fig. 5(a). (b) Q-SIAM™ preliminary map of the WV-2 T2 image shown in Fig. 8(a). Spectral categories are depicted in pseudo colors. Map legend: see Table IV. Compare this Q-SIAM™ map with the L-SIAM™ shown in Fig. 5(b), generated from the reference Landsat image depicted in Fig. 5(a).



(a) (b)

(c) (d)

Fig. 9. (a) Zoom of the WV-2 T2 image, 2 m spatial resolution, acquisition date 2010-08-04, radiometrically calibrated into TOARF values and re-calibrated according to a "master" Landsat-7 ETM+ image shown in Fig. 8(a), depicted in false colors (R: 5, G: 7, B: 2). Default image histogram stretching: ENVI linear stretching 2% [138]. (b) Zoom of the Q-SIAM™ preliminary map, shown in Fig. 6(b), of the WV-2 T2 image shown in Fig. 9(a). Spectral categories are depicted in pseudo colors. Map legend: see Table IV. (c) 4-adjacency cross-aura measure generated from the Q-SIAM™ preliminary map, shown in Fig. 9(b), of the WV-2 T1 image shown in Fig. 9(a). Cross-aura values range in $\{0, 4\}$. (d) Binary vegetation mask generated from the Q-SIAM™ preliminary map, shown in Fig. 9(b), of the WV-2 T1 image shown in Fig. 9(a).

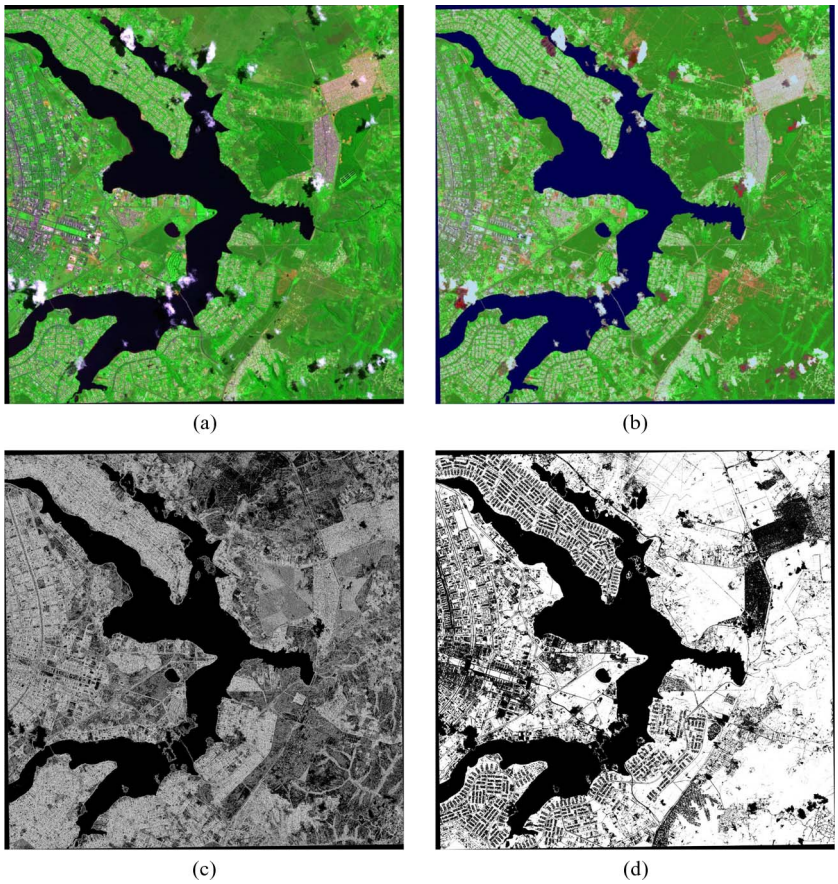


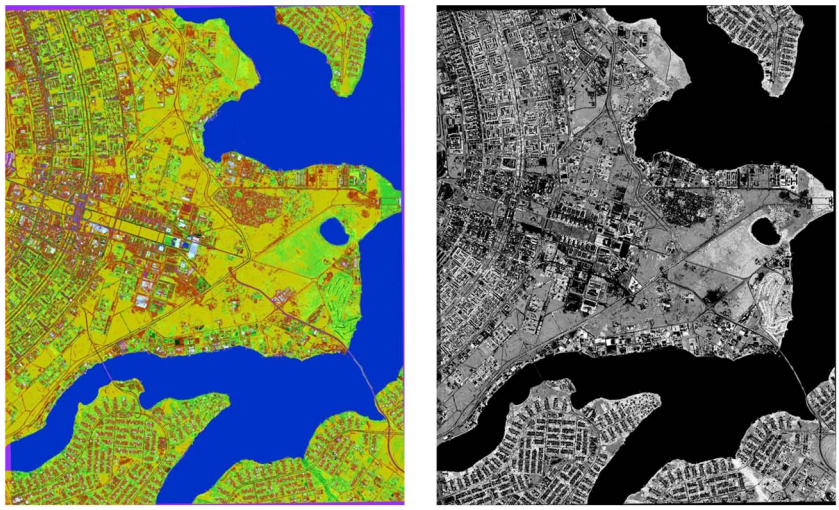
Fig. 10. (a) QB-2 $T1 + 45$ − day image, 2.4 m spatial resolution, acquisition date 2010-03-16, radiometrically calibrated into TOARF values, depicted in false colors (R: 3, G: 4, B: 1). Default image histogram stretching: ENVI linear stretching 2% [138]. (b) Q-SIAM™ preliminary map of the QB-2 $T1 + 45$ − day image shown in Fig. 10(a). Spectral categories are depicted in pseudo colors. Map legend: see Table IV. It is noteworthy that, within the Q-SIAM™ mutually exclusive and completely exhaustive classification scheme, cloud detection is $perse$ an interesting operational product with relevant commercial applications and, to the best of these authors' knowledge, without alternative solutions in either commercial or scientific RS-IUSs. (c) Four-adjacency cross-aura measure generated from the Q-SIAM™ preliminary map, shown in Fig. 10(b), of the QB-2 $T1 + 45$ − day image shown in Fig. 10(a). Cross-aura values ranges in $\{0, 8\}$. (d) Binary vegetation mask generated from the Q-SIAM™ preliminary map, shown in Fig. 10(b), of the QB-2 $T1 + 45$ − day image shown in Fig. 10(a).

It is noteworthy that the test QB-2 image was acquired approximately 45 days later than the WV-2 image at time T1, moving away from the green into the dry season. When the Q-SIAM™ preliminary classification map generated from the WV-2 T1 image, shown in Fig. 6(b), is compared with the Q-SIAM™ preliminary classification map generated from the QB-2 $T1 + 45$ days image, shown in Fig. 10(b), vegetation spectral categories typically associated with specific ranges of LAI values (refer to the Q-SIAM™ map legend shown in Table IV) reveal a moderate overall (image-wide) decrease in LAI which is perfectly consistent with seasonal effects.

This is the third source of independent evidence confirming the consistency of the radiometric registration process applied to the two WV-2 images (also refer to Section IV-A5e).

In addition, this qualitative result proves on an *a posteriori* basis that the Q-SIAM™ mapping qualities, in terms of both accuracy and robustness to changes in the input data set acquired through time and VHR sensors, namely, WV-2 and QB-2, appear high.

Since VHR optical sensors constitute a relevant portion of the available EO commercial satellite constellations, including DigitalGlobe's (which encompasses the WV-2 and QB-2 sensors investigated in this work), GeoEye's, RapidEye's, etc. (investigated in related works [5]–[17]), these conclusions mean that SIAM™ is eligible for enlarging the spectrum of large-scale (e.g., world-scale) VHR image-derived information products encompassing both thematic and continuous variables, as well as operational (near real-time, accurate, robust, easy to use, scalable) services for both scientific (quantitative) and commercial (qualitative and quantitative) RS data applications. This is tantamount to saying that commercial global providers of VHR optical images may greatly benefit (in terms of revenues, scientific impact, etc.) from the incorporation of an automatic preliminary classification first stage, like SIAM™, in operational multi-mission, multi-resolution, spaceborne/airborne optical image processing systems provided with a feedback mechanism for driven-by-knowledge RS image enhancement (e.g., automatic stratified TOC [10], refer to Sections II-G and IV-A4). For example, it is worthy of note that within the SIAM™ mutually exclusive and completely exhaustive classification scheme, cloud detection is *per se* an interesting operational product with relevant commercial applications. To the best of these authors' knowledge, the Q-SIAM™ performance in cloud detection [e.g., see Fig. 10(b)] appears superior to that of alternative commercial or scientific RS-IUSs (refer to Table I) in terms of degree of automation, accuracy and efficiency.



(a) (b)

Fig. 11. (a) Automatic SIAM™-based post-classification change/no-change detection map of the WV-2 image pair at time T1 and T2 shown in Figs. 6(a) and 8(a). Map legend: refer to Table VIII. (b) Automatic SIAM™-based post-classification change/no-change greenness (ΔGRNS) index generated from the WV-2 image pair at time T1 and T2 shown in Figs. 6(a) and 8(a), respectively. Legenda of the continuous output ΔGRNS index is as follows: 1) $\Delta\text{GRNS} < 0$ (vegetation decrease) if $(\text{GRNS}(T2) < \text{GRNS}(T1))$; 2) $\Delta\text{GRNS} > 0$ (vegetation increse) if $(\text{GRNS}(T2) > \text{GRNS}(T1))$; and 3) DumbNeg value $= (-150)$ if $((\text{GRNS}(T1) == 0)$ AND $(\text{GRNS}(T2) == 0))$, equivalent to the "never vegetation" condition.

Finally, it is worth mentioning that an operational, multi-mission, MS, spaceborne/airborne image processing feedback system incorporating SIAM™ as its preliminary classification first stage (refer to Section II-G) may be implemented:

- at the multi-mission EO image receiving antenna and ground segment, or
- on-board the satellite payload. This would open up a concrete scenario for the development of so-called fourth generation future intelligent earth observation satellites (FIEOSs) [102].

C. Q-SIAM™-Based Post-Classification Change Detection in the Test WV-2 Image Pair

It is well known that the accuracy of change/no-change detection by means of a thematic map pair difference is subjected to the following upper bound [61]:

$$\begin{aligned}&\text{Accuracy of the bi-temporal post-classification change/no-change detection map}\\ &\quad\le (\text{Accuracy of the map at time T1}\\ &\qquad\times \text{Accuracy of the map at time T2}).\end{aligned} \tag{8}$$

Thus, a map pair difference is recommended if and only if the two maps employed as input are very accurate. For example, if the two input maps feature an accuracy as high as 90%, then the accuracy of the map difference cannot be superior to 81%.

Fig. 11(a) shows the Q-SIAM™-based semantic-driven automatic bi-temporal change/no-change detection map generated from the Q-SIAM™ map pair of the two test WV-2 images at time T1 and T2 shown in Figs. 6(b) and 8(b), respectively. The legend of this automatic change/no-change detection map is shown in Table VIII.

Fig. 11(b) reveals the SIAM™-based semantic-driven automatic change/no-change greenness (GRNS) index generated from the WV-2 image pair at time T1 and T2, shown in Figs. 6(a) and 8(a), respectively, and the Q-SIAM™ map pair at time T1 and T2, shown in Figs. 6(b) and 8(b), respectively. The original GRNS index expression implemented in SIAM™ can be found in [7], [8]. Legend of the continuous output difference in greennes (ΔGRNS) is the following: 1)$\Delta\text{GRNS} < 0$ (vegetation decrease) if $(\text{GRNS}(T2) < \text{GRNS}(T1))$; 2) $\Delta\text{GRNS} > 0$ (vegetation increse) if $(\text{GRNS}(T2) > \text{GRNS}(T1))$; and 3) DumbNeg value $= (-150)$ if $((\text{GRNS}(T1) == 0)$ AND $(\text{GRNS}(T2) == 0))$, equivalent to the "never vegetation" condition.

Overall, at a qualitative level of visual assessment, the VHR difference maps shown in Fig. 11(a) and (b) appear perfectly consistent. Starting from (8), high quality of a VHR post-classification change map means that, on an *a posteriori* basis (by abduction inference [69]), the two VHR thematic maps adopted as input are both high quality.

This qualitative result is the fourth source of independent evidence confirming the consistency of the radiometric registration process applied to the two WV-2 images employed as input to the Q-SIAM™ preliminary classifier (also refer to Sections IV-A5e and IV-B).

V. Horvitz–Thompson Theorem

The objective of this section is twofold. First, the Horvitz–Thompson theorem, where unequal inclusion probabilities are accounted for probability sampling, is presented to the RS community where it is typically ignored. In this presentation, a notation partly different from that proposed in [60] is introduced. Second, original inclusion probabilities suitable for non-standard probability sampling strategies are proposed.

Map accuracy assessment is an established component of the process of creating and distributing categorical (thematic, classification) or continuous maps [55]. The fundamental basis of a categorical or continuous map accuracy assessment protocol is a location-specific comparison, across a geographic region

TABLE VIII
LEGEND OF THE AUTOMATIC Q-SIAM™-BASED POST-CLASSIFICATION BI-TEMPORAL CHANGE/NO-CHANGE DETECTION MAP

1	Constant vegetation
2	Vegetation decrease
3	Vegetation increase
4	Vegetation total gain from bare soil or built-up or fire
5	Vegetation total loss into bare soil or built-up
6	Vegetation total gain from water (or shadow)
7	Vegetation total loss into water (or shadow)
8	Single-date vegetation (affected by data noise at either T1 or T2)
9	Bare soil or built-up total gain from water (or shadow)
10	Bare soil or built-up total loss into water (or shadow)
11	Constant water (or shadow)
12	Single-date water (or shadow) (affected by data noise at either T1 or T2)
13	Constant bare soil or built-up
14	Within-bare soil or built-up change
15	Single-date bare soil (affected by data noise at either T1 or T2)
16	Constant cloud or single-date cloud (affected by noise at either T1 or T2)
17	Constant snow (or bright bare soil/built-up or cloud in VHR imagery)
18	Single-date snow (or shadowed snow) (affected by data noise at either T1 or T2)
19	Snow total gain
20	Vegetation from snow
21	Bare soil or built-up from snow
22	Water (or shadow) from snow
23	Constant shadowed snow
24	Single-date shadowed snow (affected by data noise at either T1 or T2)
25	Constant shadow
26	Constant flame
27	Single-date flame (affected by data noise at either T1 or T2)
28	Active flame
29	Constant unknown or noisy

of interest (GEOROI), between the *target map* (also called *test map* or *predicted map* [66]) to be evaluated and ground condition(s) or *"reference" condition(s)*, eventually belonging to a *complete-coverage reference map* (also called *truth map* [66]), collected from a target (true) population to be univocally identified [123].

A statistical population is defined as the collection of all discrete elements of interest together with one or more categorical and/or continuous quantities ("variables of study"), e.g., class label(s), associated with each discrete element. In map accuracy assessment, a population can be defined as all spatial units forming a partition of the GEOROI, where a spatial unit can be: 1) a point (e.g., a pixel in a digital map), 2) a polygon (e.g., a 2-D map-object), or 3) a square block of points (e.g., block of pixels), and where the categorical and/or continuous variables of study associated with each spatial unit are obtained from both the reference map and the target map [123]. In other words, both the test map and the reference map are statistical populations where each spatial unit must be associated with an instance of the categorical and/or continuous variable(s) of study. The difference between these two statistical populations is the population of interest for map accuracy assessment, where an observation of this difference population could be an indicator variable representing whether a spatial unit is classified correctly or not [123].

Unfortunately, it is impractical to obtain a census of the reference population. In other words, a complete-coverage reference map of the GEOROI almost never exists in practice. If reference conditions are available for only a sample of the GEOROI, then the test map accuracy statistics must be estimated from this reference sample [49].

There are two basic ways to approach statistical sampling (refer to Section II-E): nonprobability and probability sampling, featuring either equal or unequal inclusion probabilities. Unequal inclusion probabilities create no difficulties as long as they are known and accounted for in the estimation formulas. The inclusion probabilities determine the weight, equal to the inverse of the inclusion probability, attached to each sampling unit in the Horvitz–Thompson sample estimator. The Horvitz–Thompson theorem guarantees that the Horvitz–Thompson sample estimator is unbiased for the population total [60]. *In agreement with the QA4EO international guidelines* [3], *before being used in scientific investigations and policy decisions, thematic or continuous maps should be validated exclusively by means of probability sampling criteria* [55], [56], [60] (refer to Section II-E). Such a map accuracy assessment requirement is not obvious; for example, in the RS common practice it is traditionally violated, e.g., see [54], [61], [62].

To review the Horvitz–Thompson theorem as a unifying perspective for probability sampling, a notation partly different from that proposed in [60] is introduced as follows:

- Let U identify the finite population (universe) to be sampled. In general, the finite population U to be sampled may be intended as belonging to the test thematic map or the reference thematic map, refer to this section above.

In practice, U is a finite set of discrete population units (elements) $u \in U$, where the population size (cardinality) is $US = |U| \in \{1, \infty)$. Thus, all possible population units $u \in U$ provide a complete partition of the population U to be sampled. For example, if U belongs to a digital 2-D (image) domain, then population units belong to three spatial types: pixel, block of pixels, or polygon [123].

- S identifies the finite sample space defined as the set of all possible discrete samples (sampling units) $s \in S \equiv$ GEOROI(refer to this text above) $\supseteq U$ under the chosen probability sampling design (e.g., a simple random sampling strategy) across the GEOROI. Thus, all possible sampling units $s \in S$ provide a complete partition of the finite sample space S. For example, if $S \equiv$ GEOROI is a digital 2-D (image) domain, sampling units belong to three spatial types: pixel, block of pixels, or polygon [123].
- After sampling across the sample space $S \equiv$ GEOROI, a finite sample set $SS \subseteq S$ is selected, such that $SS \cap U = SS \subseteq U \subseteq S \equiv$ GEOROI. The SS size (cardinality) is $SSS = |SS| \in \{1, \infty)$, with $SSS \leq US$ if the element spatial types in the sample space S and universe U to be sampled are assumed to be the same.
- It is important to point out that the spatial type of samples is independent of the spatial type of population units, e.g., samples $s \in S \equiv$ GEOROI are pixels while units $u \in U \subseteq S \equiv$ GEOROI are polygons (2-D segments) or *vice versa*. For example, let us consider a finite sample space $S \equiv$ GEOROI, coincident with an EO image, where sampling units $s \in S$ are pixels whereas the target population to be sampled $U \subseteq S \equiv$ GEOROI consists of image-objects (polygons [25]) depicted in the EO image at hand and labeled as instances of class "*buildings*" by an expert photointepreter.
- $p(s)$ identifies the probability of selecting a given sample $s \in S$ at a particular step of the sampling protocol, such that [60]:

$$\sum_{s \in S} p(s) = 1, \quad \text{with } S \equiv \text{GEOROI} \supseteq U. \tag{9}$$

In a traditional sample space representation of the sampling design, a *probability sample* is defined as a selection procedure for which [60] (refer to Section I):

(i) The selection probability $p(s)$ is known for all samples $s \in S \equiv$ GEOROI $\supseteq U$, such that (9) holds true.

(ii) Each unit u in the finite population U to be sampled, with $U \subseteq S \equiv$ GEOROI, has a nonzero probability of being selected, i.e., condition $p(s) > 0$ if $(u \cap s) \neq 0$ must hold $\forall u \in U \subseteq S \equiv$ GEOROI. The first-order inclusion probability that unit u in the population U will be included in a finite sample set SS, such that $SS \cap U = SS \subseteq U \subseteq S \equiv$ GEOROI, where $SSS \leq US$ (refer to this section above), is denoted by π_{u}.

(iii) Inclusion probabilities π_{u} associated with non-sampled units $u \in U$ need only be knowable. This is extremely useful for operational needs where it is impractical to know the inclusion probabilities for the entire universe U of possible sampling units [60].

The Horvitz–Thompson estimation requires a natural transition of the sampling design representation from the selection probability representation (refer to this section above) to the inclusion probability representation. In the latter, the first-order inclusion probability π_{u} is defined for each population element (unit, object) u of the finite universe U to be sampled, whose size is US, as the probability that element u will be included in a finite sample set SS whose size is SSS, such that $SS \cap U = SS \subseteq U \subseteq S \equiv$ GEOROI with $SSS \leq US$ if the element spatial types in the sample space S and universe U to be sampled are assumed to be the same. In practice, the first-order inclusion probability π_{u} can be expressed as follows (refer to the Appendix):

$$\pi_u = \left[1 - \left(\sum_{s \in S:(u \cap s)=0} p(s)\right)^{SSS}\right] > 0, \quad \forall u \in U \subseteq S \equiv \text{ROI} \tag{10}$$

such that $\pi_{\mathrm{u}} \to 1$ if $SSS \to \infty$, i.e., unit u of the population U to be sampled is included in the finite sample set with probability tending to 1 if the sample set size tends to infinity (in compliance with the central limit theorem). In practice (10) means that the probability π_{u} of selecting a target element u in sample space S in a sequence of SSS independent yes/no experiments is equal to 1 minus the probability of selecting all the remaining non-target elements, $s \in S$: $(u \cap s) = 0$, to the power of SSS.

The inclusion probability representation of the sampling design requires that $\pi_{\mathrm{u}} > 0 \forall u \in U \subseteq S \equiv$ GEOROI. For many standard probability sampling designs, the required inclusion probabilities $\pi_{\mathrm{u}}, \forall u \in U \subseteq S$, are readily calculated. For example:

- Simple random sampling (SIRS) design. To select a random sample of n elements (sampling units) from a population U of N elements, which means that $SS \subseteq U \equiv S \equiv$ GEOROI, $SSS = n \leq US = N$, the selection probability of a sampling unit $s \in S$ is $p(s) = 1/N$, thus the probability of an element $u \in U$ of being included in the finite sample set SS becomes, according to (10):

$$\pi_u = n/N, \qquad \forall u \in SS = \{s_1, \ldots, s_n\}. \tag{11}$$

  For further details about (11), refer to the Appendix.
- Stratified random sampling (STRS) design, where strata $h = 1, \ldots, H$ are available *a priori*, such that their size in terms of elements (sampling units) N_h is known, e.g., the number of sampling units n_h must be increased to reduce the standard error (confidence interval) of class-specific accuracy estimates such that the stratum h corresponds to the mapped area of the h-th class [123]. If selection within stratum h of n_h sampling units out of N_h elements known a priori is conducted via a stratum-specific SIRS, which means that $SS_h \subseteq U_h \equiv S_h \equiv \text{GEOROI}_h$, $SSS_h = n_h \leq US_h = N_h$, then the selection probability of a sampling unit $s \in S_h$ is $p(s) = 1/N_h$, thus the probability of an element $u \in U_h$ of being included in the finite sample set SS_h becomes, according to (10) as a generalization of (11):

$$\pi_u = n_h/N_h, \qquad \forall u \in SS_h = \{s_1, \ldots, s_{n_h}\}. \tag{12}$$

As a non-standard sampling example, let us conceive a probability sampling design capable of generating a random sample of the finite population U of (2-D) image-objects, labeled as *buildings*, which are depicted in a EO image whose total area is $A =$ number of rows $\times$ number of columns. Suppose we randomly locate 20 sample points (s_pnts) across the spaceborne image, such that GEOROI(image) $\equiv S$, to generate a sample set $SS = \{s\_pnt_1, \ldots, s\_pnt_{20}\}$, whose cardinality is $SSS = 20$. The selection probability $p(s)$ would be the probability that sample point $s \in$ sample space$S$ "hits" a target image-object (polygon) u. This is tantamount to estimating the probability $p(s)$ of selecting a sample polygon (s_plygn), $s \in S \equiv$ GEOROI $\supseteq U$, where s coincides with a target image-object (building) $u \in U$. If a sample object (s_plgyn) $s \equiv u$ has area au and the area of the sample space $S \equiv$ GEOROI (image) is A, then the random selection probability for a sample polygon $s\_plygn \equiv u$ is

$$p(s\_plygn \equiv u) = \sum_{i=1}^{au} p(s\_pnt_i) = \sum_{i=1}^{au} 1/A = au/A,$$
$$s\_pnt_i \in S \equiv \text{GEOROI} \supseteq U.$$

Hence, according to (10), the probability that at least one of the 20 sample points "hits" an image-object $u \in U$ is

$$\pi_u = \left\{1 - \left\lfloor p(s\_plygn \neq u)^{SSS} \right\rfloor\right\} = \left\{1 - [(A - au)/A]^{20}\right\}.$$

The same sampling strategy described above implemented within a given stratum U_h whose area is A_h provides an inclusion probability equal to

$$\pi_u = \left\{1 - \left\lfloor p(s\_plygn \neq u)_h^{SSS} \right\rfloor\right\} = \left\{1 - \lfloor (A_h - au)/A_h \rfloor^{20}\right\}.$$

The Horvitz–Thompson theorem may be stated as follows [73]: if $\pi_u = (10) > 0$, $\forall u \in U \subseteq S \equiv$ GEOROI, then, for a given finite sample set SS, such that $SS \cap U = SS \subseteq U \subseteq S \equiv$ GEOROI, with cardinality $|SS| = SSS > 0$, the Horvitz–Thompson sample estimator

$$\hat{T}_y = \sum_{u \in SS} \frac{y_u}{\pi_u} = \sum_{u \in SS} w_u \cdot y_u \tag{13}$$

where

$$w_u = 1/\pi_u, \qquad \forall u \in SS \subseteq U \subseteq S \equiv \text{GEOROI} \tag{14}$$

is unbiased for the population total [60],

$$T_y = \sum_{u \in U} y_u \tag{15}$$

i.e., (13)$\rightarrow$(15) must hold, where (14) means that each sampled population element $u \in SS \subseteq U \subseteq S \equiv$ GEOROI, represents w_u elements of the finite population U (which has been sampled) when the sample data statistics are "expanded" to estimate totals and means over U.

For example, in a standard STRS design, if there are two strata (populations to be sampled independently) U_1 and U_2 with, respectively, $N_1 = 1000$, $N_2 = 100$, and $n_1 = n_2 = 20$, then $\pi_u = (12) = 20/1000 = 1/50$ in U_1 and $\pi_u = (12) = 20/100 = 1/5$ in U_2, thus each sample element $u \in SS_1 \subseteq U_1 \subseteq S_1 \equiv \text{GEOROI}_1$ represents $w_u = 50$ elements of the population U_1 being sampled, whereas each sample element $u \in SS_2 \subseteq U_2 \subseteq S_2 \equiv \text{GEOROI}_2$ represents $w_u = 5$ elements of the population U_2 being sampled.

As another example, if N is the number of elements (units) in the finite population (universe) U to be sampled according to a standard SIRS strategy eligible for selecting n sampling units when $y_u = 1$, then $T_y = (15) = N$ while $\pi_u = (11) = n/N$, $\forall u \in SS \subseteq U \subseteq S \equiv$ GEOROI, thus:

$$\hat{T}_y = \sum_{u \in SS} \frac{y_u}{\pi_u} = \sum_{u=s_1}^{s_n} w_u \cdot 1 = \sum_{u=s_1}^{s_n} \frac{N}{n} = N = T_y$$

hence, (13) $=$ (15) holds in line with theoretical expectations, which means that (11) $= \pi_u = n/N$ is correct.

Analogously, in a standard STRS design, if N_h is the number of elements (units) in the h-th finite stratum U_h to be sampled to select n_h sampling units when $y_u = 1$, then $T_y = (15) = N_h$ while $\pi_u = (12) = n_h/N_h, \forall u \in SS_h \subseteq U_h \subseteq S \equiv$ GEOROI, thus:

$$\hat{T}_y = \sum_{u \in SS_h} \frac{y_u}{\pi_u} = \sum_{u=s_{1,h}}^{s_{n,h}} w_u \cdot 1 = \sum_{u=s_{1,h}}^{s_{n,h}} \frac{N_h}{n_h} = N_h = T_y$$

hence (13) $=$ (15) holds in line with theoretical expectations, which means that (12) $= \pi_u = n_h/N_h$ is correct.

To recapitulate, because of the emphasis on equal probability sampling in introductory statistical methods courses, statistics practitioners may view unequal probability sampling as unacceptable or even "biased". The Horvitz–Thompson theorem establishes the intuitively appealing solution that unequal inclusion probabilities of sampled population elements, $u \in U \subseteq S \equiv$ GEOROI, are accounted for simply by using the appropriate per-unit weight w_u equal to the inverse of the inclusion probability, $w_u = (1/\pi_u) = (14)$[60]. For many standard probability sampling designs, like SIRS and STRS, the required inclusion probabilities $\pi_u, \forall u \in U \subseteq S$, are readily calculated, see (11) and (12), otherwise they need to be carefully addressed (refer to this section above).

VI. Novel Protocol to Operationalize the Thematic and Spatial Accuracy Assessment of Thematic Maps Generated From VHR Spaceborne/Airborne Imagery

In this section, the six components of a novel probability sampling protocol for accuracy assessment of thematic maps generated from VHR imagery are discussed and instantiated in compliance with the probability sample design proposed by Stehman and Czaplewski [55] (see Section I). These six stages are summarized below.

(i) *Identification of the GEOROI, test map taxonomy, reference sample set taxonomy and their contingency table*.

(ii) *Probability sampling design*, where the following decisions must be taken.

- Estimation of the sample set cardinality depending on the project's requirements specification in terms

of: 1) target OA or per-class accuracies, 2) target confidence interval, and 3) available project budget.
- Selection of the sampling frame. The sampling units providing a complete partition of the sampling universe can be represented by a sampling frame [123]. A sampling frame consists of the materials or devices which delimit, identify, and allow access to the elements of the target population across the GEOROI. There are two types of sampling frames: (1-D, 1-D) list frames and (2-D, 2-D) area frames [55].
- Selection of the spatial type(s) of sampling units, e.g., pixel, polygon or block of pixels [123].
- Selection of the sampling strategy, e.g., SIRS (refer to Section V), systematic sampling, STRS (refer to Section V), etc.

(iii) *Evaluation protocol*, namely, procedures to collect information pertaining to the thematic determination of both reference and test sampling units. Typically, information pertaining to the thematic determination of the reference sampling units is collected by means of field campaigns, photointerpretation of EO images "one step closer to the ground" than the RS data used to make up the test map [51], i.e., EO images whose spatial and/or spectral quality is higher than that of the RS images employed for the generation of the test map, or a combination of these two information sources.

(iv) *Labeling protocol*, consisting of rules for assigning one or more class indexes to each reference sampling unit and each test sampling unit based on the information collected by the evaluation protocol.

(v) *Analysis protocol*, where a contingency table (error matrix) selected in step (i) is instantiated.

(vi) *Estimation protocol*, where summary QIs, provided with their confidence interval, are estimated from the contingency table(s) and assessed in comparison with reference standards in compliance with the QA4EO international guidelines [3] (refer to Section II-D).

In the rest of this section, the aforementioned six stages are designed and instantiated for accuracy assessment of preliminary classification maps automatically generated at multiple semantic granularities by the automatic SIAM™ software product (refer to Section II-G) from three VHR test images acquired across time, space, and sensors (refer to Section IV).

A. Identification of the GEOROI, Test Map Taxonomy, Reference Sample Set Taxonomy, and Their Contingency Table

A thematic map accuracy assessment begins with the identification of the GEOROI (refer to Section V), test map taxonomy (legend), reference sample set taxonomy, and their contingency table (error matrix).

Whenever multiple sensory data or information sources [23] must be combined (integrated) and used together in a structured (synergistic) way, "as there may be differences in semantics as well as in the structure of these data sets, the data must be adapted to fit the task, often with compromises being made. Currently, the cost of these integration and adaptation activities is a major barrier to the adoption and efficient exploitation of complex data sets. An important aspect of this integration process is the recognition of semantic differences between data sets. Often these differences are missed due to incomplete documentation, but more importantly mistakes occur because of misunderstanding due to assumptions made at the domain level... Subtle differences in semantics may result in data being improperly integrated, which may not be noticed until after operational decisions are made... These mistakes may be costly" [70].

In practice, the development of ontologies (e.g., spatio-temporal ontologies of the world-through-time, refer to Section II-C) may facilitate the capture of domain knowledge in such a way as to detect or prevent errors when semantic data sources must be integrated. In the words of philosophical hermeneutics [23], [24], where an inquirer (receiver, cognitive agent) always plays a pro-active role in the generation of information from data, the concept of (qualitative) "*information-as-(an interpretation) process*" is complementary to the concept of (quantitative) "*information-as-thing*" adopted by the Shannon theory of communication [78]. In the notion of "*information-as-(an interpretation) process*," a "*fusion of horizons*," or "*fusion of ontologies*," always takes place between a speaker and the listener(s) (refer to Section II-A). For example, in the artificial intelligence common practice, an application-domain expert (knowledge expert) provides a set of requirements, in user-speak [124], for the content and scope of a conceptual ontology, to be later transformed by a knowledge engineer [69] into a statement of external functionality, in techno-speak [124], suitable for developing the logical ontology expressed in description logics [70].

In the domain of thematic map accuracy assessment and comparison, two semantic data sources, the test map and the reference sample set, must be compared. Thus, the semantic data source integration (harmonization) process must begin with the identification and clear understanding of:

(i) the legend (taxonomy [68], ontology [70], [80] or classification system [51], which includes both taxonomy and generation rules [51], [68]) of the test map to be evaluated across the selected GEOROI,

(ii) the reference classification extracted from the target, finite, and discrete population located in the GEOROI and

(iii) relations required to harmonize the test and reference semantic vocabularies. In the words of Cerba *et al.* [140], "harmonisation of classifications schemes and systems, codelists, terminology and vocabulary (i.e., selection of corresponding items, definition of rules for mapping languages) must be created before the building of (data) harmonisation tools".

The problem of inter-vocabulary semantic mapping is clearly acknowledged by the existing literature where thematic maps comparison is considered a fundamental procedure in geographical analysis [66]. For example, Stehman describes four common types of categorical map comparison [63].

1) Comparison of different thematic maps, either crisp of fuzzy [64], [65], of the same region and featuring the

same sorted set of LC classes, which is tantamount to saying different categorical maps of the same GEOROI with the same thematic map legend, taxonomy [68], ontology [70], [80] or semantic vocabulary, considered as a sorted set of semantic concepts. Typical motivations to compare maps that share the same categorical variable for the same region are [66]: 1) to quantify landscape transformation by comparing a map from a former time to a map from a latter time, 2) to validate a simulation model by comparing a *predicted map* to a *truth map*, and 3) to evaluate cartographic techniques by comparing a map created by one technique to a map created by an alternative technique. To date, a large segment of the RS community appears concerned with this first type of maps comparison exclusively [51], [66].

2) Comparison of thematic maps, either crisp of fuzzy, of the same region, but whose LC class vocabularies differ in terms of semantics and/or order of presentation and/or cardinality (number of classes) [68], [139], [140]. This second type of thematic map comparisons includes the first type as a special case.
3) Comparison of thematic maps, either crisp of fuzzy, of different regions, but featuring the same map legend.
4) Comparison of thematic maps, either crisp of fuzzy, of different regions and whose map legends differ in terms of semantics and/or order of presentation and/or cardinality. This fourth type of thematic map comparisons includes the third type as a special case.

About the comparison of thematic map pairs, the following considerations hold:

- Although a large segment of the RS community appears exclusively concerned with the first of the aforementioned four types of comparison of thematic maps, there is a significant body of literature dealing with the second and/or fourth type. In the words of Ahlqvist [68, p. 1227]: "many scholars have acknowledged a need to negotiate and compare information stemming from different classification systems. Works on semantic uncertainty [125] and semantic interoperability [126] of geographic information reflect this concern. Works in computer science, artificial intelligence and information science have also tackled the issue of translations between heterogeneous information sources and many see a potential for using formalized descriptions, or ontologies [70], that can describe category semantics, to address this issue [71]–[73]. . . Once a classification scheme has been transformed into a formalized categorization a translation can be achieved by matching the concepts in one system with concepts in another, either directly or through an intermediate classification. Conceptually, these computational approaches largely follow suggestions from the cognitive sciences on how categories are mentally constructed and this has informed some recent examples of negotiating different nomenclatures including some that specifically target incompatible land use and LC taxonomies [74]–[76]".
- Semantic associations between the test and reference semantic vocabularies involved with a categorical map pair comparison, equivalent to inter-vocabulary semantic relations considered as "*correct*" by a cognitive agent (refer to Section II-A), are, in general, many-to-many relations [139], whose special (simpler) cases are one-to-many, many-to-one, and one-to-one relations.
- All possible semantic associations between the two discrete and finite sets of test and reference semantic concepts are represented by the Cartesian product (or product set) between the two discrete and finite sets of concepts. In a tabular form, a Cartesian product is equivalent to a so-called bi-dimensional *contingency table* [55], otherwise called *association matrix*, *cross-tabulation matrix* [66], *full semantic change matrix* [68], *error matrix* [55], or OAMTRX [52], [67], which can be either square or non-square, depending on whether cardinalities of the two nominal sets are equal or different.
- The matching (harmonization) of two semantic legends is, *per se*, a cognitive (interpretation) problem whose solution is equivocal, pertaining to the domain of "*information-as-(an interpretation) process*" (refer to Section II-A). It means that two independent cognitive agents (knowledge engineers in the nomenclature of artificial intelligence [69]) are likely to match two thematic vocabularies differently. In other words, *no "universal (context-independent) best match" between two categorical variables can exist, but the most appropriate semantic match between different nomenclatures becomes a matter of negotiation and community-agreement* [23], [24].
- A comprehensive interpretation of an OAMTRX, either square or non-square, can be very challenging, complex, and time consuming. In general, an OAMTRX has no major diagonal of matching class_pairs ("*correct*" table entries) to guide the interpretation [52], [67], [68]. This is tantamount to saying that off-diagonal or scattered table entries can be considered perfectly "*correct*" in an either square or non-square OAMTRX.
- A square OAMTRX is not equivalent to a popular CMTRX, which is square by definition [49], [50], [51]. In the former, the presence of "*correct*" off-diagonal or scattered entries is expected. In the latter, the test and reference sorted legends coincide and the main diagonal guides the interpretation process. A square OAMTRX becomes equivalent to a traditional (square) CMTRX if and only if the test and reference semantic vocabularies are the same sorted set of concepts. In other words, it is always true that the class of CMTRXs is a subset (special case) of the class of OAMTRXs, i.e., OAMTRX $\supset$ CMTRX.
- It is noteworthy that whereas the construction of an OAMTRX is straightforward and non-controversial when the semantic labels of sampling units are crisp (hard), the method to construct an OAMTRX is not obvious at all when semantic labels are soft (fuzzy) [66], [68].
- A similar consideration holds about the selection of quality summary statistics generated from an OAMTRX. Although there is still some debate on this issue when dealing with crisp semantic labels [49], [127]–[129], there is even more debate when fuzzy semantic labels are involved [66], [68].

This work adopts the following hypotheses:

- It exclusively deals with the first and second types of the aforementioned four types of thematic map comparisons, where the second type is a generalization of the first (refer to this section above).
- It exclusively deals with hard semantic labeling of each sampled population element belonging to the test map domain or the reference sample.

Based on these working hypotheses, the following considerations hold true:

- "*Correct*" associations between the test and reference semantic vocabularies are expected to be many-to-many. They can be modeled as "*correct*" entries in an OAMTRX, which can be either square or non-square.
- Since it deals with crisp semantic labels exclusively, the construction of an OAMTRX becomes straightforward and non-controversial (refer to this section above).
- Although an OAMTRX has no major diagonal of "*correct*" entries to guide the interpretation [52], [67], [68], the distribution of "*correct*" entries across an OAMTRX does convey (useful) information about the degree of match between the two test and reference categorical variables [23]. Intuitively, the information carried out by many-to-many relations encompassed by an either square or non-square OAMTRX is: 1) more vague (fuzzier) than, i.e., inferior to, say, the unambiguous information carried out by ideal (simplest) one-to-one relations typically allowed by a CMTRX, and 2) superior to the null information conveyed by all-to-all inter-vocabulary relations. To estimate from an OAMTRX the degree of match between the two test and reference categorical variables, the novel Categorical Variable Pair Similarity Index, $CVPSI \in [0,1]$, is proposed below.

1) Definition of the Novel CVPSI Estimated From an OAMTRX: An original degree of match between a test and a reference categorical variable, identified as $CVPSI \in [0,1]$, is estimated from an OAMTRX as described below.

In an OAMTRX (and its special case, CMTRX), it is typical that columns represent the reference classification while rows indicate the test map to be evaluated [51]. Let us identify as TC the cardinality of the test classification taxonomy and as RC the cardinality of the reference sample taxonomy. The total number of "*correct*" (allowed) elements (cells, entries) in an OAMTRX is identified as CE, such that $0 \le CE \le RC \times TC$. In addition, symbol "==" is adopted to mean "equal to."

The CVPSI computation problem is constrained as follows:

(a)
$$CE = \sum_{t=1}^{TC}\sum_{r=1}^{RC} CE_{t,r}, \text{with } CE_{t,r} \in \{0,1\} = \{\text{"correct" entry}(t,r), \text{"noncorrect" entry}(t,r)\}, \quad CE \in \{0, RC \times TC\}.$$

(b) If $(CE == 0)$, then $CVPSI = 0$. It means that, when no "*correct*" entry exists, then the degree of match between the two categorical variables is zero. If $(CE == 0)$, then also the classification OA probability estimate, $p_{\mathrm{OA}} \in [0,1]$ is equal to 0.

(c) If $(CE == RC \times TC)$, then $CVPSI \to 0$. It means that when all table entries are considered "*correct*," then nothing is meaningful or makes the difference between the two categorical variables. It is noteworthy that, if $(CE == RC \times TC)$, then $p_{\mathrm{OA}} = 1$, but at the expense of $CVPSI \to 0$. This intuitively proves that, being independent QIs of the test thematic map under consideration, p_{OA} and $CPVSI$ must be maximized jointly.

(d) If
$$\left\{\left[\left(\sum_{t=1}^{TC} CE_{t,r} = CE_{+,r}\right) == 1, r = 1, \ldots, RC\right] \times AND \left[\left(\sum_{r=1}^{RC} CE_{t,r} = CE_{t,+}\right) == 1, t = 1, \ldots, TC\right]\right\}$$
i.e., if $[(CE_{RC} = RC)$ AND $(CE_{TC} = TC)]$, then $CVPSI = 1$. It means that when the reference and test map legends "match" each other by means of one-to-one relations exclusively, then the OAMTRX is equivalent to a CMTRX and CVPSI is maximum.

(e) If [*not* condition(b) *AND not* condition(c) *AND not* condition(d)] then $CVPSI \in (0,1)$.

For example, in a (square) CMTRX, then $CVPSI = 1$ according to condition (d). In practice, $CVPSI \in [0,1]$ is a fuzzy degree of similarity between: 1) an OAMTRX whose definition requires the selection by a domain expert of the "*correct*" entries, i.e., "*correct*" (allowed) reference-test class relations which are, in general, many-to-many and 2) an (ideal) CMTRX version of an OAMTRX, where allowed reference-test class relations are one-to-one exclusively, irrespective of the fact that "*correct*" entries are diagonal or off-diagonal entries.

Another way of interpreting index $CVPSI \in [0,1]$ is to consider its complementary value $(1 - CVPSI) \in [0,1]$. Index $(1 - CVPSI)$ is a *normalized estimate of the additional (classification) work required to fill up the semantic gap from the test semantic vocabulary to the reference semantic vocabulary*.

To satisfy the set of aforementioned constraints (a) to (e), the following set of original equations is proposed.

$$CVPSI \in [0,1], CVPSI = \frac{1}{RC+TC}\left(\sum_{r=1}^{RC} f_{RC}(CE_{+,r}) + \sum_{t=1}^{TC} f_{TC}(CE_{t,+})\right) \quad (16)$$

with

$$f_{RC}(i) = \begin{cases} 0 & \text{if } i = 0, \\ e^{-\frac{(i-1)^2}{\left(\frac{TC}{3}\right)^2}} & \text{if } i > 0, \end{cases} \quad \text{with } i \in \{0, TC\} \subset I_0^+, \text{ where } i = CE_{+,r}, r \in \{1, RC\} \quad (17)$$

$$f_{TC}(j) = \begin{cases} 0 & \text{if } j = 0, \\ e^{-\frac{(j-1)^2}{\left(\frac{RC}{3}\right)^2}} & \text{if } j > 0, \end{cases} \quad \text{with } j \in \{0, RC\} \subset I_0^+, \text{ where } j = CE_{t,+}, t \in \{1, TC\}. \quad (18)$$

It is trivial to prove that (16)–(18) satisfy the aforementioned requirements (a) to (d). In this section below, it is proved that requirement (e) is satisfied too.

TABLE IX
REFERENCE LAND COVER CLASS DESCRIPTIONS AND SAMPLING UNIT TYPES IDENTIFIED IN THE WV-2 AND QB-2 TEST IMAGES DESCRIBED IN SECTION IV AND CONSIDERED AS THE REFERENCE POPULATION TO SAMPLE

Reference Class	Spatial type of sample units [123]	Description
Dark Built-up (*DB*)	Polygon	Manmade foreground features distinguished by low to average spectral response in visible wavelengths. Examples: asphalt roads; parking lots; dark brown, gray, or black buildings/ roofs
Light Built-up (*LB*)	Polygon	Manmade foreground features distinguished by high spectral response in visible wavelengths.Examples: white/tan concrete or cement roads; white/tan concrete or cement parking lots; white/tan buildings/roofs
Tree Crown (*TCrwn*)	Polygon	Foreground vegetation distinguished from other background vegetation (e.g. grass) by vertical attribute and, typically, low to medium spectral response in near-infrared wavelengths. Examples: individual trees; individual shrub
Bare Soil (*BS*)	Pixel	Background (ground-level) land cover consisting of non-vegetated, barren land. Examples: exposed rocks; light or dark soil
Grass (*Gr*)	Pixel	Background (ground-level) land cover consisting of vegetation typically displaying medium to very high spectral response in near-infrared wavelengths. Examples: manicured lawns; pastures; grasslands; rangelands
Shadow in general or Cloud Shadow or Cloud (*ShC*)	Pixel	Background or foreground features which include strong shadows or clouds (clouds present only in QuickBird-2 image).Examples: strong shadow; thin cloud cover; thick cloud cover
Water (*Wa*)	Pixel	Background water features displaying low spectral response at all wavelengths. Examples: deep clear water; shallow clear water; turbid water

TABLE X
DEFINITION OF "CORRECT" ENTRIES IN AN OAMTRX INSTANTIATION [8], [52], [67] GENERATED AS THE CROSS-TABULATION BETWEEN THE Q-SIAM™ SPECTRAL CATEGORIES AT THE INTERMEDIATE LEVEL OF SEMANTIC GRANULARITY [5]–[17], WHERE $TC = 28$ (SEE TABLE V), AND THE ADOPTED REFERENCE SET OF LAND COVER CLASSES, WHERE $RC = 7$ (4 TABLE IX). THUS, IN ITS COMPLETE VERSION, THIS CONTINGENCY TABLE CONSISTS OF $RC = 7$COLUMNS AND $TC = 28$ ROWS. FOR THE SAKE OF SIMPLICITY, IN THIS EXAMPLE (SYNTHETIC, BUT REALISTIC), ONLY 16 OF THE 28 SIAM™ SPECTRAL CATEGORIES, NAMELY, THOSE WHOSE OCCURRENCE IS ABOVE 0.5% IN THE TEST MAP GENERATED FROM THE WV-2 TEST IMAGE, IN ADDITION TO CLASS "*Unknowns*" ($UN3$), ARE SHOWN. THE YELLOW HIGHLIGHT COLOR IDENTIFIES "*Correct*" ENTRIES (ALLOWED PAIRWISE RELATIONS). THE PINK HIGHLIGHT COLOR IDENTIFIES BARRED ENTRIES. IN THIS OAMTRX EXAMPLE, IF RC IS FIXED TO 7 AND TC IS SET EQUAL TO 16, THEN THE INTER-VOCABULARY DEGREE OF MATCH $CVPSI = 0.5598$ (REFER TO THE TEXT IN THIS SECTION FOR COMPUTATION DETAILS)

		Reference Classes						
SIAM™ super-categories, pseudo-color	**SIAM™ non-leaf (SC) and leaf spectral categories (LSC), pseudo-colors and acronyms (refer to [6])**	Dark Built-up (*DB*)	Light Built-up (*LB*)	Tree Crown (*TCrwn*)	Bare Soil (*BS*)	Grass (*Gr*)	Cloud shadow or cloud or shadow (*ShC*)	Water (*Wa*)
Vegetation	SV_SC	×	×	✓	×	✓	×	×
	AV_SC	×	×	✓	×	✓	×	×
	SHRWE_LSC	×	×	✓	×	×	✓	×
	SHV_WEDR_LSC	×	×	✓	×	×	✓	×
	ASHRBR_SC	×	×	✓	×	✓	×	×
	AHRBCR_LSC	×	×	✓	×	✓	×	×
	PB_LSC	×	×	✓	×	✓	×	×
Bare soil or built-up	BBB_VBBB_SC	✓	✓	×	✓	×	✓	×
	SADBBVF_SC	✓	✓	×	✓	×	×	×
	SADBBF_SC	✓	✓	×	✓	×	×	×
	SADBBNF_SC	✓	✓	×	✓	×	×	×
	SHB_LSC	×	×	×	✓	×	✓	×
Water or shadow	DPWASH_LSC	×	×	×	×	×	✓	✓
	TWASH_LSC	✓	✓	×	×	×	✓	✓
Snow or cloud or bright bare soil	SN_CL_BBB_LSC	✓	✓	×	✓	×	✓	×
Unknowns	UN3_LSC	×	×	×	×	×	×	×

2) Instantiation With the SIAM™ Pre-Classification Maps Automatically Generated from the WV-2 and QB-2 Test Images: SIAM™'s preliminary classification maps consist of a discrete and finite set of spectral categories at fine, intermediate, and coarse semantic granularities (refer to Section II-G) [5]–[17]. Many-to-many associations hold between SIAM™'s *color-based inference categories* (e.g., "*vegetation*"), which belong to the (2-D) image domain, and a reference set of LC classes (e.g., "*deciduous forest*," "*evergreen forest*," etc.), equivalent to 4-D object-models belonging to the 4-D model of the world-through-time [25], [35] (refer to Section II-C). Hence, an either square or non-square OAMTRX is required to model the many-to-many relations capable of harmonizing the test spectral categories and the reference LC classes [8], [52], [67] (refer to this section above).

In general, when dealing with a map generated from VHR imagery, no data source one step closer to the ground than the VHR image employed to make up the map is available for reference population sampling. Hence, the test and reference VHR data sources coincide [51] (refer to Section III). This is

also the case of this work, where a reference LC taxonomy is selected by an expert photointerpreter (cognitive agent), whose inference activities are, *per se*, subjective (equivocal [23], [24], refer to Section II-A), in the VHR test image at hand according to the following constraints:

(a) The reference legend is mutually exclusive and totally exhaustive, in compliance with the Congalton and Green criteria for selecting a target taxonomy [51].
(b) It is capable of mapping reference image-objects, identified by the expert photointerpreter in the test VHR image at hand, into a discrete and finite set of semantic labels.
(c) The reference legend and the reference image-objects are selected by independent means from the RS-IUS whose maps, generated from the same VHR test image, are being validated (refer to Section III).

As a result of a photointerpretation process of the WV-2 and QB-2 test images subjected to the aforementioned constraints by a domain expert, seven reference LC classes are identified as shown in Table IX.

As an example, Table X shows the OAMTRX instance [8], [52], [67] whose "*correct*" entries are defined for the Q-SIAM™ spectral categories at the intermediate level of semantic granularity [5]–[17], where $TC = 28$, (see Table V) cross-tabulated with the reference set of LC classes, listed in Table IX, where $RC = 7$. Thus, in its complete version, this contingency table consists of $RC = 7$ columns and $TC = 28$ rows. To reduce the vertical size of Table X, only 16 of the 28 Q-SIAM™ spectral categories, namely, those whose occurrence is above 0.5% in the test map generated from the WV-2 image, in addition to class "*Unknowns*" (*UN3*), are shown.

It is noteworthy that, in Table X, selection of "*correct*" entries is equivalent to the "*fusion of horizons*," or "fusion of ontologies," mentioned above in this section. Although subjective in nature (because terms "*semantic*," "*subjective*," and "*equivocal*" are synonyms [23], [24], refer to Section II-A), *associations between spectral categories and reference LC classes shown in Table X should not be considered arbitrary. Rather, they should be community-agreed and considered specific of the two test and reference semantic vocabularies, but independent of the RS image adopted as input by the test thematic map*.

As a synthetic but realistic example, let us investigate the inter-vocabulary degree of match, CVPSI, estimated, via (16)–(18), from the OAMTRX instance shown in Table X when RC is fixed to 7 and TC is set equal to 16.

- If all elements in Table X were considered as "*correct*" entries, thus $CE = RC \times TC = 7 \times 16 = 112$, then condition (c) would hold, where it is expected that $CVPSI \to 0$. If we apply (16)–(18) we obtain:

Equations (16) to (18)

$$= CVPSI = \frac{1}{7+16}\left(7 * f_{RC}(16) + 16 * f_{TC}(7)\right)$$

$$= \frac{1}{22}(7 * 0.000367 + 16 * 0.001344) = 0.001 \approx 0.$$

Thus, (16)–(18) satisfy constraint (c) when an OAMTRX features all "*correct*" entries, when $RC = 7$ and $TC = 16$.

- In the specific case of Table X, where $RC = 7$, $TC = 16$ and $CE = 42 \leq RC \times TC = 7 \times 16 = 112$, then requirement (e) should hold, therefore:

Equations (16) to (18)

$$= \frac{1}{7+16}\left(\sum_{r=1}^{RC=7} f_{RC}(AE_{+,r}) + \sum_{t=1}^{TC=16} f_{TC}(AE_{t,+})\right)$$

$$= \frac{1}{23}\left(f_{RC}(2) + f_{RC}(5) + 3 * f_{RC}(6) + 2 * f_{RC}(7) + \sum_{t=1}^{16} f_{TC}(AE_{t,+})\right)$$

$$= \frac{1}{23}\left(0.965455 + 0.569783 + 3 * 0.415237 + 2 * 0.282063 + \sum_{t=1}^{TC} f_{TC}(CE_{+t})\right)$$

$$= \frac{1}{23}\left(3.345074 + f_{TC}(0) + f_{TC}(1) + 8 * f_{TC}(2) + 3 * f_{TC}(3) + 2 * f_{TC}(4) + f_{TC}(5)\right)$$

$$= \frac{1}{23}\left(3.345074 + 0 + 1 + 8 * 0.832208 + 3 * 0.479652 + 2 * 0.191463 + 0.052931\right)$$

$$= \frac{1}{23}(3.345074 + 9.532473) = \frac{12.87747}{23}$$

$$= 0.559893 \in (0, 1).$$

The conclusion is that (16)–(18) satisfy constraint (e) in the example where OAMTRX = Table 1.

Table XI reports the CVPSI value extracted from the three OAMTRX instances defined in this work to cross-tabulate the reference set of LC classes, shown in Table IX, where $RC = 7$, with the legend of a Q-SIAM™ map at fine, intermediate and coarse semantic granularity, where the number of Q-SIAM™'s test classes, TC, is equal to 52, 28, and 12, respectively, refer to Table V. For the sake of simplicity, these three OAMTRX instances are not shown in this presentation, but can be accessed through anonymous ftp [148]. In addition, a subset of the OAMTRX instance defined between the Q-SIAM™ legend at intermediate granularity and the reference classes is shown in Table X.

In line with theoretical expectations, Table XI reveals that, in these experiments, the CVPSI index is monotonically non-decreasing (i.e., it increases or remains constant) with the semantic granularity of the Q-SIAM™ map, i.e., CVPSI does not increase as the SIAM™ semantic granularity gets coarser. In other words, in these experiments, CVPSI is monotonically non-decreasing with the degree of specialization of the Q-SIAM™ spectral categories. *Vice versa*, the "*semantic gap*" from the test Q-SIAM™ spectral categories to the reference land cover classes, estimated as $(1 - CVPSI)$, is monotonically non-increasing (i.e., it decreases or remains constant) with the cardinality of the latter [5]–[17].

TABLE XI
CATEGORICAL VARIABLE PAIR SIMILARITY INDEX, $CVPSI \in [0,1]$, REFER TO (16)–(18), ESTIMATED FROM AN OAMTRX INSTANCE WHERE "*Correct*" ENTRIES ARE SELECTED BETWEEN THE Q-SIAM™ SPECTRAL CATEGORIES, WHOSE CARDINALITY TC AT FINE, INTERMEDIATE, AND COARSE SEMANTIC GRANULARITY IS EQUAL TO 52, 28, AND 12, RESPECTIVELY, AND THE REFERENCE CLASSES, DESCRIBED IN TABLE IX, WHOSE CARDINALITY $RC = 7$. HENCE, THESE THREE OAMTRX INSTANCES CONSIST OF 364, 196, AND 84 ENTRIES. FOR THE SAKE OF SIMPLICITY, THESE THREE OAMTRX DEFINITIONS ARE NOT SHOWN IN THIS PRESENTATION, BUT CAN BE ACCESSED THROUGH ANONYMOUS FTP [148]. IN ADDITION, A SUBSET OF THE OAMTRX INSTANCE DEFINED BETWEEN THE Q-SIAM™ LEGEND AT INTERMEDIATE GRANULARITY AND THE REFERENCE CLASSES IS SHOWN IN TABLE X.

Test map legend's cardinality = Q-SIAM™ semantic granularity (refer to Table 5)	***CVPSI***
Fine, $TC = 52$. The number of reference classes $RC = 7$, see Table IX.	0.6715
Intermediate, $TC = 28$. The number of reference classes $RC = 7$, see Table IX.	0.6134
Coarse, $TC = 12$. The number of reference classes $RC = 7$, see Table IX.	0.5136

Finally, it is noteworthy that the CVPSI estimation from an OAMTRX definition, like that shown in Table X, is preliminary to and completely independent from the instantiation phase of the OAMTRX, when cells are filled with occurrences or probability values. These probability values are investigated by TQIs, like the OA probability. This is to say that, being independent QIs of the test thematic map under investigation, TQIs and the CPVSI must be maximized jointly.

B. Probability Sampling Design

In this subsection, a probability sampling protocol is proposed to accomplish the following decisions (refer to the introduction to Section VI).

- Estimation of the sample set cardinality depending on the project's requirements specification, with regard to: 1) target OA and confidence interval, 2) target per-class accuracy and confidence interval, and 3) costs of sampling in compliance with the project budget.
- Selection of the sampling frame, either (1-D) list frame or (2-D) area frame [55].
- Selection of the spatial type(s) of sampling units, e.g., pixel, polygon, or block of pixels [123].
- Selection of the sampling strategy, e.g., SIRS, STRS, systematic sampling, etc. (refer to Section V).

1) Sample Set Cardinality Estimation: In order to estimate the minimum number of reference sampling units to be sampled and labeled for each thematic class of the test map to be evaluated, Lunetta and Elvidge propose a statistical criterion which depends on the project requirements specification, namely, the target class-specific accuracy and error tolerance, but is independent of costs of sampling to be accounted for in the project budget [50]. This criterion is described below.

It is well known that any classification OA probability estimate, $p_{\mathrm{OA}} \in [0,1]$, is a random variable (sample statistic) with a confidence interval (error tolerance) associated with it, identified as $\pm\delta$, where δ represents the half-width of the error tolerance at a specified *confidence level*$(1-\alpha)$ such that $0 < \delta < p_{\mathrm{OA}} \leq 1$, with $\alpha \in [0,1]$, known as the desired *level of significance* (e.g., $\alpha = 0.05$), which is the risk that the actual error is larger than $\pm\delta$, hence the specified confidence level $(1-\alpha)$ (e.g., $1-\alpha = 1-0.05 = 95\%$) is the required probability that the actual error falls within the confidence interval $\pm\delta$. In practice, $p_{\mathrm{OA}} \pm \delta$ is a function of the specific test data set used for its estimation, and *vice versa*. For example, for a given reference sample set size (SSS) comprising independent and identically distributed (i.i.d.) reference samples[2] and an estimated classification accuracy probability p_{OA}, it is possible to prove that the half width δ of the error tolerance $\pm\delta$ at a desired confidence level (e.g., if confidence level $(1-\alpha) = 95\%$ then the critical value is 1.96) can be computed as follows [50]:

$$\delta = \sqrt{\frac{(1.96)^2 \cdot p_{OA} \cdot (1-p_{OA})}{SSS}}. \tag{19}$$

Vice versa, minimum $SSS = f$ (target p_{OA}, target δ) can be computed as follows:

$$SSS = \frac{(1.96)^2 \cdot p_{OA} \cdot (1-p_{OA})}{\delta^2}. \tag{20}$$

For each c-th class simultaneously involved in the classification process, with $c = 1, \ldots, C$, where C is the total number of classes, with $C \geq 2$ (at least, the total number of classes C comprises a target LC class and class "*outliers*"; It is noteworthy that the definition of a rejection rate is a well-known objective of any RS image classification system, e.g., refer to [94, p. 185], it is possible to prove that [50]):

$$\delta_c = \sqrt{\frac{\chi^2_{(1,1-\alpha/C)} \cdot p_{OA,c} \cdot (1-p_{OA,c})}{SSS_c}}, c = 1, \ldots, C \tag{21}$$

where α is the desired level of significance, i.e., the risk that the actual error is larger than $\pm\delta_c$ (e.g., $\alpha = 0.05$), $1-\alpha/C$ is the level of confidence (e.g., if $\alpha = 0.05$ and $C = 5$, then $1 - 0.05/5 = 0.99$), and $\chi^2_{(1,1-\alpha/C)}$ is the upper $(1-(\alpha/C))^*$ $100th$ percentile of the chi-square distribution with one degree of freedom (e.g., if the level of confidence is $(1 - 0.05/5) = 0.99$, then $\chi^2_{(1,0.99)} = 6.63$).

Vice versa, minimum $SSS_c = f$ (target $p_{\mathrm{OA,c}}$, target δ_{c}), $c = 1, \ldots, C$, can be computed as follows:

$$SSS_c = \frac{\chi^2_{(1,1-\alpha/C)} \cdot p_{OA,c} \cdot (1-p_{OA,c})}{\delta_c^2}, c = 1, \ldots, C. \tag{22}$$

Instantiations of community-agreed target values and confidence intervals of classification accuracy measures can be

[2]In the RS common practice, the i.i.d. hypothesis almost never applies to reference samples due to spatial autocorrelation between neighboring pixels belonging to the same LC type. This is tantamount to saying that the number of statistically independent observations that can be made in space-time is limited [25]. This is in accordance with Tobler's first law of geography: "all things are related, but nearby things are more related than distant things" [130], although certain phenomena clearly constitute exceptions [25].

found in literature. For example, according to the USGS classification system constraints [111], the target one-class $p_{\mathrm{OA}} \in [0,1] \pm \delta$ is fixed at $0.85 \pm 2\%$. The per class classification accuracy, $p_{\mathrm{OA,c}} \in [0,1] \pm \delta_{\mathrm{c}},\ c = 1, \ldots, C$ should be about equal and never below 70%, whereas a reasonable reference standard for δ_{c} is about 5% [8].

This means that, if the desired level of significance $\alpha = 0.07$ and $C = RC = 7$, then the level of confidence $(1 - \alpha/C) = 0.99$ and $\chi^2(1, 1 - \alpha/C) = 6.63$. In this case, if $p_{OA,c} = 85\%$, with $\delta_{\mathrm{c}} = \pm 2\%$, then $SSS_c = (22) = 2113$, $c = 1, \ldots, RC = 7$. If $p_{OA,c} = 85\%$, with $\delta_{\mathrm{c}} = \pm 5\%$, then $SSS_c = (22) = 338$, $c = 1, \ldots, RC = 7$, and so on.

Instantiation with the SIAM™ pre-classification maps automatically generated from the WV-2 and QB-2 test images: In this paper, the project requirements specification is as follows:

- The target number of reference LC classes, RC, is set equal to $RC = 7$, see Table IX.
- The target OA probability and confidence interval, $p_{\mathrm{OA}} \in [0,1] \pm \delta$, are fixed at $0.85 \pm 2\%$, in line with the USGS standards [111]. The significance level, $\propto$, is fixed at 0.05, thus $\chi^2_{(1,1-\propto)} = \chi^2_{(1,1-0.05)} \approx 3.84$ in (20).
- Per-class accuracies, $p_{OA,c}$, $c = 1, \ldots, C$, with C equal to either $TC = 52/28/12$ (see Table V) or $RC = 7$, should be similar and never below 0.70, with an error tolerance, $\pm\delta_{\mathrm{c}}$, equal to $\pm 5\%$. For the estimation of the reference per-class sample size $SSS_c = (22)$, the target $p_{OA,c}$, $c = 1, \ldots, RC$, is set equal to 0.85 [8]. In addition, the reference per-class significance level, $\propto /RC$, is fixed at 0.01, thus $\chi^2_{(1,1-\propto/RC)} = \chi^2_{(1,1-0.01)} \approx 6.63$ in (22).

Given these project requirements, sample set size estimates are calculated as follows:

- According to (20), the minimum sample set size, independent of the test image and sampling costs, necessary to assess the OA assuming USGS parameters is

$$SSS = (20) = \frac{\chi^2_{(1,1-\alpha)} \cdot p_{OA} \cdot (1 - p_{OA})}{\delta^2} \approx \frac{3.84 \cdot 0.85 \cdot (1 - 0.85)}{0.02^2} \approx 1225.$$

- According to (22), the minimum per-class sample set size necessary to assess the target per-class accuracy assuming the previously defined parameters is

$$SSS_c = (22) = \frac{\chi^2_{(1,1-\alpha/RC)} \cdot p_{OA,c} \cdot (1 - p_{OA,c})}{\delta_c^2},$$

$$c = 1, \ldots, RC \approx \frac{6.63 \cdot 0.85 \cdot (1 - 0.85)}{0.05^2} \approx 340.$$

It is noteworthy that these minimum sample set size estimates may refer to sample units whose spatial type is either pixel or polygon, e.g., refer to Table V.

2) Sampling Frame: The definition of a sampling design requires to specify a finite sample space S, assumed to coincide with the selected GEOROI, i.e., $S \equiv$ GEOROI, see Section VI-A, where S consists of a finite set of discrete spatial units (sampling units, e.g., pixels, blocks of pixels, or polygons [123]) that form a complete (totally exhaustive) partition of the selected GEOROI, such that S is a superset of the finite population U to be sampled, thus $U \subseteq S \equiv$ GEOROI. The sampling units forming the 2-D sampling universe $S \equiv$ GEOROI can be represented by a sampling frame [123]. There are two types of sampling frames: a *(1-D. 1-D) list frame* and a *(2-D, 2-D) area frame* [55].

- A 1-D list frame is defined as a list of all such spatial units forming a complete partition of the target GEOROI along with a spatial address for each unit (e.g., spatial coordinates or an identification number unique to each unit). Thus, the sample is selected directly from this 1-D list of sampling units representing the entire GEOROI independent of the 2-D sample space S [123].
- The sampling protocol used with an area frame is based on, first, selecting a sample of ideal dimensionless *spatial locations*, also called *sample candidates* or *sampled locations* [55], otherwise called *geo-atoms* as a dimensionless atomic abstraction of geographic information [25], followed by associating a sampling unit with each spatial location. Hence, the actual sampling units, for example, polygons, are selected indirectly via the intermediate step of the sample of point locations. An explicit rule for associating a unique sampling unit, say, either a pixel, polygon, or block of pixels, with any spatial location within the area frame must be established. For example, a rule for associating a unique polygon with a randomly selected point location is to sample that polygon within which the random point fell. This particular area frame sampling protocol illustrates that it is not necessary to delineate all polygons in the target population to obtain the sample. An area frame is preferable to a 1-D list frame when a systematic design is planned. For example, if the area frame is a map of all pixels, converting the map to a 1-D list frame of pixels would not only be unnecessary work, it would lose much of the spatial structure important for systematic sampling. Area frames better retain the spatial features of a geospatial population [55]. On the other hand, point (area) sampling is effectively the same as sampling from a 1-D list frame if the spatial units (e.g., pixels) partitioning the GEOROI are all equal in area and have the same shape. For example, for a polygon assessment unit, the point sampling protocol will select polygons into the sample with probability proportional to polygon area, so larger polygons will have a higher probability of being selected. These unequal inclusion probabilities must be accounted for in the analysis and estimation stage, e.g., by means of the Horvitz–Thompson theorem, see Section V.

Instantiation with the SIAM™ pre-classification maps automatically generated from the WV-2 and QB-2 test images: In this paper, neither a complete-coverage reference map ("*truth map*" [66]) is available nor reference categorical strata (layers) are available on a *a priori* basis. For example, no reference map of the image-wide set of image-objects identified as elements of the reference class "*Light Built-up*" (LB), see Table IX, is available. Thus, no 1-D list frame is possible. An area frame is selected instead, such that if the sampling unit is an areal

Fig. 12. Original non-standard class-specific simple random sampling (SIRS) strategy. Left. Sample point selection for the reference class "*Grass*" (Gr), see Table IX, using a set of random spatial locations (sample candidates) generated by a SIRS strategy applied per reference class. Since the reference class Gr features non-contextual (i.e., color) properties, rather than contextual properties, like geometric attributes or texture, then the selected spatial unit is "pixel," see Table IX. Green locations ("*hit*") are included in the Gr class-specific sample, while red points ("*miss*") are excluded. Right. Sample polygon selection for the reference class "*Light Built-up*" (LB), see Table IX, using a set of random spatial locations. "Hits" on an image-object which is a member of the reference class LB require delineation of the object-specific shape and size with a polygon, illustrated with a red boundary. This LB class-specific object selection shows how a single reference class-object consists of different colors likely to be matched by several spectral categories detected by SIAM™, i.e., this semantic image-object belonging to reference class LB is eligible for encompassing a one-to-many relation with spectral categories (spectral end members) detected by the first-stage SIAM™ preliminary classifier.

sampling unit, e.g., sample pixel or polygon, then the explicit rule for associating a sample pixel or polygon with a randomly selected point location is to select that sample pixel or polygon within which the random point fell, see Fig. 12.

3) Selection of the Spatial Type(s) of Sampling Units: The sampling unit, e.g., a 0.1 hectare (ha) pixel, 10 ha polygon, 1000 ha circular plot, etc., is the fundamental unit on which the accuracy assessment is based. The sampling unit can be defined without specifying what will be observed on that unit on the ground; thus, no assumption about homogeneity of thematic classes for the sampling unit is necessary. Because the sampling unit is the ultimate basis for the comparison of the thematic map and reference sample classifications, whatever sampling unit is chosen, it is essential that this choice be explicitly and clearly stated and considered acceptable to users of the thematic map [55].

There are two types of sampling units [55].

- (Dimensionless) Point, featuring no area extent. The statistical population associated with a point sampling unit is viewed as continuous. In [25], dimensionless atomic abstraction of geographic information are called *geo-atoms* (refer to this section above).
- Areal unit featuring a 2-D spatial coverage. The statistical population associated with areal units is considered as partitioned into discrete spatial units such as pixels or polygons. The three primary areal sampling units are the following.
  - Pixels, representing small areas (e.g., 30 m pixel), are related to point sampling units, but because pixels still possess some areal extent, they partition the mapped population into a finite, though large, number of sampling units.
  - Polygons. Polygon sampling units are usually irregular in shape and differ in size. For example, in a RS imagery, man-made objects are typically photointerpreted as polygons approximating the object-specific shape and size, see Fig. 12.
  - Fixed-area plots. Fixed-area plot sampling units are usually regular in shape and cover some predetermined areal extent. In practice, pixels and polygons are special cases of fixed-area plot sampling units.

Instantiation with the SIAM™ pre-classification maps automatically generated from the WV-2 and QB-2 test images: In the case of Q-SIAM™ maps generated from the WV-2 and QB-2 test images described in Section IV, areal sampling units, either pixels or polygons, are selected class specific, refer to Table IX.

4) Selection of the Probability Sampling Strategy: SIRS, STRS (equivalent to running an SIRS of n_h elements from the N_h elements in stratum h), systematic sampling (with a random start and sampling interval K, with K an integer), and cluster sampling are all probability sampling designs considered as reference standards because they guarantee that (refer to Section V): 1) each element u in the population U to be sampled has a positive inclusion probability, $\pi_{\mathrm{u}} > 0$, $\forall u \in U$, 2) the probability of an element $u \in U$ being included in an arbitrary sample set SS of the population U, with $SS \subseteq U \subseteq S \equiv$ GEOROI, where the sample space $S \equiv$ GEOROI, is known, e.g., see equations. (11) and (12), and 3) inclusion probabilities associated with non-sampled units need only be knowable [60].

About STRS, it may be important to recall that geospatial (e.g., categorical) strata must be available before (prior to) the geospatial statistical sampling takes place, i.e., STRS is possible if and only if prior geospatial knowledge in the form of strata is available before the statistical sampling occurs.

Instantiation with the SIAM™ pre-classification maps automatically generated from the WV-2 and QB-2 test images: In our experiments, neither a complete-coverage reference map nor reference class-specific strata are available. For example, in Table IX the reference class LB features a spatial type equal to polygon, but no LB class-specific (stratum-specific) reference map is available, i.e., there is no prior knowledge of where polygons identifying units of reference class LB are located across the VHR image to be sampled for accuracy assessment of the thematic map generated from that VHR image (refer to Section VI-A).

Thus, *no standard STRS is possible since no categorical stratum is available a priori. As a consequence, to implement an area frame* according to Section VI-B2, *an original nonstandard class-specific SIRS strategy is applied as described in the caption of* Fig. 12. For example, in Fig. 12, where a sample polygon selection for the reference class LB (refer to Table IX) is implemented by using a set of random spatial locations, "hits" on an image-object which is a member of the reference class LB require manual delineation of the target object-specific shape and size with a polygon, illustrated with a red boundary, to be drawn by an expert photointerpreter. In practice, since no 1-D list of reference image-objects (polygons) belonging to the reference class LB is available, a human photointerpreter is required to select a finite set of reference class-specific image-objects as those "hit" by a (theoretically infinite) set of spatial random locations until the required cardinality of the reference sample set, e.g., estimated via (22), is accomplished.

To deal with the unequal inclusion probability of reference polygons in compliance with the Horvitz–Thompson theorem (see Section V), let us consider the probability sampling strategy described above. Unfortunately, it differs from standard probability sampling designs, such as SIRS and STRS, for which the required inclusion probabilities $\pi_{\mathrm{u}}, \forall u \in U \subseteq S$, are readily calculated, e.g., see (11) and (12). In addition, it deals with a multi-class classification problem where class-specific strata across a target 2-D GEOROI (image) are not known, therefore the inclusion probability derived from (10), $\pi_{\mathrm{u}} = 1 - [p(s\_plygn \neq u)]^{SSSh} = 1 - [(A_h - au)/A_h]^{SSSh}$ proposed in Section V, where SSS_h is the sample set size for stratum h, au is the area of the 2-D object u belonging to the population (stratum) U_h to be sampled and A_h is the area of stratum U_h, cannot be applied because parameter A_h is unknown.

To recapitulate, *a novel inclusion probability of reference polygons in compliance with the Horvitz–Thompson theorem (see* Section V) *must be introduced in the probability sampling strategy adopted in this work*. As an adaptation of (10) proposed in Section V, for each sample polygon s_plygn belonging to the sample space, i.e., $s_plygn \in S \equiv$ GEOROI, coincident with a target class-specific image-object $u_c \in U_c \subseteq S \equiv$ GEOROI with $c \in \{1, RC\}$, where the class-specific cardinality of the finite sample set SS_{c} for class c, $|SS_{\mathrm{c}}| = SSS_{\mathrm{c}}$ is defined according to (22), the original inclusion probability $\pi_{\mathrm{u,c}}$ adopted in this work is proposed as follows:

$$\pi_{u,c} = \frac{area_{u,c}}{\sum_{n=1}^{SSS_c} area_{n,c}} \in (0,1], \quad area_{u,c} \geq 1 \text{ (in pixel units)}, \quad \text{where } \sum_{n=1}^{SSS_c} \pi_{n,c} = 1. \quad (23)$$

C. Response Phase Protocol

According to the introduction to Section VI, the response phase consists of an *evaluation protocol* and a *labeling protocol.*

- The *evaluation protocol* specifies how the reference information will be collected and integrated from different sources of reference ("true") classification.
- The *labeling protocol* includes rules for assigning one or more (e.g., a primary and a secondary) reference classifications to each sampling unit.

1) Evaluation Protocol: The *evaluation protocol* specifies how the reference information will be collected and integrated from different sources of reference ("true") classification such as, say, photointerpretation of high-quality RS imagery (e.g., VHR imagery), field campaigns, or a combination of these information sources. In practice, the evaluation protocol starts from choosing the size and shape of the *spatial support region (domain of activation)* where the reference sample classification evaluation will occur based on evidences collected from a selected set of reference information sources such as VHR imagery, field campaigns, or both. In particular:

- if the sampling unit is a dimensionless point (geo-atom [25]), the evaluation need not be limited only to what the evaluator observes at that point location.
- If the sampling unit is an areal sampling unit, namely, pixel, polygon, or blocks of pixels [123], a spatial support region defined for an areal sampling unit may or may not be the areal unit itself. For example, a 30 m pixel may be assigned with a support region of 1 ha. The spatial support of a polygon sampling unit will usually just be the polygon itself [55].

Instantiation with the SIAM™ pre-classification maps automatically generated from the WV-2 and QB-2 test images: In these experiments, where the sole reference information source available for sampling is the same VHR image adopted as input by Q-SIAM™ to make up the thematic maps to be evaluated (refer to Section III), seven mutually exclusive and totally exhaustive reference classes, together with their spatial units, are identified in the test VHR image set, see Table IX. The spatial support of a polygon sampling unit is considered the polygon itself plus its spatial context, say, a square block of pixels 10 times the size of the sampled object (e.g., in a VHR image the identification of a polygon as an instance of LC class "*buildings*" may be reinforced if in its surrounding there is a road linked with that image-object). The spatial support of a pixel sampling unit is considered a 20 × 20 block of pixels (e.g., equal to 40 × 40 m in a WV-2 image) centered on the sampled pixel.

2) Labeling Protocol: It consists of rules for assigning one or more reference class indexes to each sampling unit based on the information obtained from the evaluation protocol. It may provide: 1) crisp labeling, where more cover types can be selected if it is not possible or desirable to label the sampling unit as a single thematic class, or 2) fuzzy labeling, where a class membership is provided for every class in the reference taxonomy.

Instantiation with the SIAM™ pre-classification maps automatically generated from the WV-2 and QB-2 test images: In these experiments, where the sole reference information source available for sampling is the same VHR image adopted as input by Q-SIAM™ to make up the thematic maps to be evaluated, a crisp labeling strategy is adopted by an expert photointerpreter (cognitive agent), whose (inherently equivocal) mapping means are required to be independent of the RS-IUS that generates the maps to be evaluated, refer to Section VI-A.

D. Analysis and Estimation Protocol

To cope with the well-known *non-injectivity of any QI*, no hypothetical universal QI exists [5]–[17]. Hence, the estimation and analysis protocol for a statistically rigorous quality assessment of a thematic map in comparison with reference geospatial sampling units must rely on a selected set of mutually uncorrelated QIs provided with uncertainty in measurement in compliance with the QA4EO guidelines [3] (refer to Section III).

The level of detail provided by spaceborne/airborne VHR images allows observation as discernible image features (image-contours or, *vice versa*, image-objects) of ground level (3-D) objects (e.g., cars) traditionally invisible in coarser resolution spaceborne images, whose typical areal spatial type is pixel rather than polygon. The ill-fated dichotomy of pixels versus image-objects (polygons) remains on open issue to cope with in the development of operational RS-IUSs [5]–[17], [39], [40], [77], [79], [99], [131] (refer to Section II-F). The same dichotomy between pixels and image-objects affects the accuracy assessment of thematic maps generated from spaceborne/airborne VHR images. Traditional thematic maps made up from low- and medium-resolution RS images require a pixel-based assessment of the map's thematic and spatial accuracies [52]. In addition to a pixel-based thematic accuracy assessment, categorical maps generated from VHR imagery require a spatial (geometric) accuracy assessment of the map polygons, which accounts for the spatial distribution of thematic errors [52], [53], [56] (refer to Section III). To summarize, unlike a traditional thematic accuracy assessment, which is pixel-based, spatial accuracy assessment, mandatory for maps generated from VHR images, is 2-D object based. In particular, the goal of a spatial accuracy assessment of a thematic map is to investigate: 1) the precision of a reference object's boundary delineation through scale and 2) the appropriateness of a reference object's area and shape through scale [79].

Hence, two sets of symbolic pixel-based TQIs and subsymbolic polygon-based SQIs are proposed hereafter, starting from existing literature and in compliance with the QA4EO guidelines (refer to Section III).

1) Thematic Accuracy Assessment: In line with recommendations found in a relevant portion of the existing literature, where the use of popular pixel-based TQIs, such as the *kappa* coefficient, is strongly discouraged [49], [127]–[129], pixel-based TQIs selected in this paper are the traditional OA probability (p_{OA}), user's accuracy (p_{usr}) and producer's accuracy (p_{prdcr}). These measures directly illustrate the probability of encountering a correct or incorrect labeled pixel, i.e., they allow comparisons between digital maps consisting of map units $u \in U$ as pixels [49], irrespective of the spatial type of sample units $s \in S$ as either polygon (s_plygn) or pixel (s_pxl), where the sample space $S \equiv \text{GEOROI} \supseteq U$ (refer to Section V).

In the specific case of a (square) CMTRX of size $C \times C$, where C is the total number of LC classes, the test and reference semantic vocabularies coincide, columns usually represent the reference classification while rows indicate the test map to be evaluated [49]–[51], then the p_{OA} index, defined as the sum of the main diagonal elements (correctly classified pixels), is computed as:

$$p_{OA} = \sum_{c=1}^{C} p_{c,c}. \tag{24}$$

However, presenting the sole p_{OA} is not enough, because different thematic maps may feature the same non-injective p_{OA} index value (refer to Section III). Every error is an omission from the correct category and a commission to a wrong category [51]. Producer's and user's accuracies are ways of representing individual category accuracies instead of just the overall classification accuracy. User's accuracy $p_{usr,c}$ represents the conditional probability that an area classified as c, $c = 1, \ldots, C$, by the test map is also classified as class c by the reference sample. Thus, user's accuracy $p_{usr,c}$, $c = 1, \ldots, C$, is related to the inverse of the commission error (false positive). In the specialized case of a (square) CMTRX, user's accuracy is computed as follows [49], [51]

$$p_{usr,c} = \frac{p_{c,c}}{\sum_{i=1}^{C} p_{c,i}} = \frac{p_{c,c}}{p_{c,+}} \tag{25}$$

where $p_{c,+}$ is the row total (row marginal). Similarly, producer's accuracy $p_{prdcr,c}$ represents the conditional probability that an area classified as c by the reference sample is also classified as class c by the test map. Thus, producer's accuracy $p_{prdcr,c}$, $c = 1, \ldots, C$, is related to the inverse of the omission error (false negative). It is computed as:

$$p_{prdcr,c} = \frac{p_{c,c}}{\sum_{i=1}^{C} p_{i,c}} = \frac{p_{c,c}}{p_{+,c}} \tag{26}$$

where $p_{+,c}$ is the column total (column marginal).

In the more general case of a (square or non-square) OAMTRX $\supset$ CMTRX [67], it is necessary to adjust the numerators of (24)–(26), originally formulated to suit (square) CMTRX instances where one-to-one associations between the test and reference semantic vocabularies hold, to fit

many-to-many relations, which are typical of OAMTRX instances, whose special cases are relations many-to-one, one-to-many, and one-to-one, e.g., refer to the "*correct*" entries in Table X.

Instantiation with the SIAM™ pre-classification maps automatically generated from the WV-2 and QB-2 test images: Equations (24)–(26) are implemented to deal with the "*correct*" entries identified, in accordance with Section VI-A, in the nine OAMTRX instances estimated from the three-granule Q-SIAM™ maps (see Table V) generated from the three VHR test images (refer to Table VI). For the sake of simplicity, the nine estimated OAMTRX instances are not shown in this presentation, but they can be accessed through anonymous ftp [148].

Hence, nine instantiations of the TQI (24)–(26) are computed, where the cardinality of the test semantic vocabulary, TC, is equal to 52/28/12 for the Q-SIAM™ maps at fine, intermediate and coarse semantic granularity, respectively (see Table V), while the cardinality of the reference semantic vocabulary, RC, is equal to 7 (refer to Table IX). Collected TQI values are shown in Table XII(a)–(c) for each of the three VHR test images. At first glance, Table XII(a)–(c) show that the estimated overall, user's and producer's accuracies tend to far exceed the target mapping accuracy values specified in Section VI-B1. For example, OA p_{OA} is greater than or very close to 99% for the three VHR test images across the three semantic granularities.

When confidence intervals are taken into account, the overall accuracies $p_{OA} \pm \delta$ reported in Table XII are found to overlap significantly across sensors, acquisition dates, and semantic granularity, although the WV-2 T1 classification maps at the three levels of semantic granularity show the best (by a statistically irrelevant tiny bit) OA p_{OA} among the three test images. These best cases are consistently associated with the same input image, which may be considered as an additional evidence of the consistency of the Q-SIAM™ maps generated at various semantic granularities from the same VHR image.

With rare exceptions which are examined later in this section, user's accuracy estimates, $p_{usr,c}$, $c = 1, \ldots, TC = 52/28/12$, surpass the target 70% specified in Section VI-B1 and, in most cases, they exceed 95%, while producer's accuracy estimates, $p_{prdcr,c}$, $c = 1, \ldots, RC = 7$, are equal or above their minimum at 94.49%, and greater than 98% in all but two cases.

In terms of user's accuracy, an exception showing low accuracy (high commission error) is spectral category SHV_WEDR ("*shadow vegetation or weak dark rangeland*") in the WV-2 T2 image-derived maps at fine and intermediate semantic granularities, see Tables XII(a) and (c), where this spectral category features a high degree of commission with asphalt surfaces. This effect is likely the result of undesired saturation introduced through the relative calibration procedure in the WV-2 T2 image (refer to Section IV). In the QB-2 image, the spectral categories ASHRBR_VLNIR ("*average shrub rangeland with very low NIR*") and SBBVF ("*strong barren land or built-up with very flat spectral response*") also display low users's accuracy (high commission error) in the fine granularity classification. These errors stem from the presence of thin clouds, which are absent from the two WV-2 image acquisitions.

In the QB-2 image, both user's and producer's accuracy show a slight decline in some classes in comparison with the pair of WV-2 data-derived maps, which is due to presence of thin clouds causing mixed pixel effects. When moving from fine granularity to the intermediate semantic granularity level, misclassification stemming from clouds and mixed pixels becomes less apparent, e.g., producer's accuracy (inversely related to omission errors) for the reference class "*Shadow or Cloud Shadow or Cloud*" (*ShC*, see Table IX) remains approximately the same while user's accuracies (inversely related to commission errors) increase, from minimum values of $48.02 \pm 3.39\%$ (ASHRBR_VLNIR) and $72.04 \pm 3.78\%$ (SBBVF) in Table XII(a), to minimum values of $90.57 \pm 0.30\%$ (SADBBVF, "*strong or average or dark barren land or built-up with very flat spectral response*") and $94.26 \pm 9.24\%$ (SHB, "*shadow area with barren land*") in Table XII(b).

To properly interpret the "high" *TQI values reported in* Tables XII(a)–(c), *it is of fundamental importance to consider that these TQIs are derived from an OAMTRX instance, like that shown in* Table X, *where many-to-many relations between two different test and reference semantic vocabularies are considered as "correct"* (refer to Section VI-A). These "*correct*" many-to-many relations account for a "*semantic degree of match*" between the two test and reference semantic vocabularies (semantic horizons) estimated as $CVPSI \in [0, 1]$, see (16)–(18), or, *vice versa*, $(1 - CVPSI) \in [0, 1]$, which is the semantic distance between the two legends proportional to the additional (classification) work required to fill up the semantic gap from the test to the reference semantic vocabulary. This means that, in these experiments, TQIs are expected to be somehow "high" due to the semantic vagueness inherent with the harmonization of two different test and reference semantic vocabularies (refer to Section VI-A). For example, the SIAM™ spectral categories have a semantic meaning superior to zero, but equal or below that of reference LC classes, refer to Section II-G. *Since their semantic content is "vague" (equivalent to a high level of abstraction or low level of specialization), the SIAM™ spectral categories are capable of generalization at the cost of specialization. In practice, TQI values reported in* Tables XII(a)–(c) *show that, since spectral categories are broad concepts, they are almost never wrong*. Experimental evidences supporting this conceptual reasoning are found in Tables XII where, for each of the three test images, the Q-SIAM™ map generated at coarse semantic granularity features: 1) the highest (by a statistically irrelevant tiny amount) OA across different granularities and 2) the highest $(1 - CVPSI)$ value of the semantic gap from the test to the reference taxonomy. This means that, *in the proposed experiments, a coarser semantic granularity of the Q-SIAM™ map achieves a (slightly) higher thematic accuracy at the expense of semantic specialization (information utility, informative content)*.

In addition, to interpret correctly the TQI values shown in Tables XII(a)–(c), it is important to recall here that SIAM™ is an automatic deductive (prior knowledge-based) decision-tree classifier (expert system, refer to Section II-B) requiring no training phase before running the mapping stage (refer to Section II-G). In other words, collected TQIs reported in Tables XII(a)–(c) cannot be positively biased by any high

TABLE XII

(a) OVERALL, USER'S, AND PRODUCER'S ACCURACY FOR THE SIAM™ FINE SEMANTIC GRANULARITY CLASSIFICATION OF TEST DATA SETS REPRESENTED IN PERCENTAGES. SYMBOL *—DENOTES ABSENCE OF TARGET POPULATION OR SMALL TARGET POPULATION (PRESENCE IN IMAGE $< 0.5\%$). SYMBOL ^—ERROR LIKELY DUE TO EFFECTS OF RELATIVE CALIBRATION PROCEDURE OR SENSOR SATURATION

		QuickBird-2		WorldView-2, T1		WorldView2, T2	
Q-SIAM™ super-categories at fine semantic granularity	**Overall Accuracy, $p_{OA}\pm\delta$: refer to equations (19) and (24), with $\alpha = 0.05$, $(1-\alpha) = 0.95$. Inter-vocabulary degree of match *CVPSI* = 0.6715, with *TC* = 52, refer to Table 11.**	98.97 ± 0.34		99.57 ± 0.24		99.44 ± 0.24	
Q-SIAM™ super-categories at fine semantic granularity, pseudo-colors	**SIAM™ leaf spectral categories (LSC, at fine semantic granularity), pseudo-colors and acronyms (in parentheses are the *Reference Classes* according to Table 10), User's Accuracy, $p_{usr,c}\pm\delta_c$, c = 1, ..., *TC* = 52: refer to equations (21) and (25), with α = TC/100, 1- α/TC = 0.99, χ^2 = 6.63.**	**QuickBird-2**		**WorldView-2, T1**		**WorldView2, T2**	
Vegetation	SVVH2NIR_LSC (*TCrwn, Gr*)	100.00±0.00		0.00±0.00*		100.00±0.00*	
	SVVH1NIR_LSC (*TCrwn, Gr*)	100.00±0.00		100.00±0.00		100.00±0.00	
	SVVHNIR_LSC (*TCrwn, Gr*)	100.00±0.00		100.00±0.00		99.94±0.17	
	SVHNIR_LSC (*TCrwn, Gr*)	100.00±0.00		100.00±0.00		100.00±0.00	
	SVMNIR_LSC (*TCrwn, Gr*)	100.00±0.00		100.00±0.00		100.00±0.00	
	SVLNIR_LSC (*TCrwn, Gr*)	100.00±0.00*		0.00±0.00*		100.00±0.00*	
	AVVH1NIR_LSC (*TCrwn, Gr,ShC*)	100.00±0.00		0.00±0.00*		96.32±11.43*	
	AVVHNIR_LSC (*TCrwn, Gr, ShC*)	100.00±0.00		100.00±0.00		98.98±1.15	
	SHRWE_LSC (*TCrwn, ShC*)	99.65±0.94		100.00±0.00		99.90±0.26	
	SHV_WEDR_LSC (*TCrwn, ShC*)	93.84±3.98		98.70±1.05		50.74±3.00^	
	ASHRBRHNIR_LSC (*TCrwn, Gr*)	100.00±0.00		100.00±0.00		99.51±0.19	
	ASHRBRMNIR_LSC (*TCrwn, Gr*)	99.35±0.15		98.57±0.17		99.42±0.13	
	ASHRBR LNIR_LSC (*TCrwn, Gr*)	99.98±0.04		99.99±0.02		98.00±0.26	
	ASHRBR VLNIR_LSC (*TCrwn, Gr*)	48.02±3.39		80.66±2.61		98.28±0.39	
	AHRBCR_LSC (*TCrwn, Gr*)	0.00±0.00*		0.00±0.00*		100.00±0.00*	
	PB_LSC (*TCrwn, Gr*)	97.35±0.81		94.24±4.74		97.57±0.88	
	GH_CL_LSC (*TCrwn, Gr, ShC*)	98.58±13.26		0.00±0.00*		0.00±0.00*	
Bare soil or built up	VBBB_TNCL_LSC (*LB, BS, ShC*)	99.68±0.93		100.00±0.00*		100.00±0.00	
	BBB_TNCL_LSC (*LB, BS, ShC*)	100.00±0.00		100.00±0.00		100.00±0.00	
	SBBVF_LSC (*DB, LB, BS*)	72.04±3.78		100.00±0.00		100.00±0.00	
	SBBF_LSC (*DB, LB, BS*)	100.00±0.00		100.00±0.00		100.00±0.00	
	SBBNF_LSC (*DB, LB, BS*)	99.61±0.33		100.00±0.00		100.00±0.00	
	ABBVF_LSC (*DB, LB, BS*)	87.60±0.57		100.00±0.00		100.00±0.00	
	ABBF_LSC (*DB, LB, BS*)	100.00±0.00		100.00±0.00		100.00±0.00	
	ABBNF_LSC (*DB, LB, BS*)	98.66±0.33		100.00±0.00		100.00±0.00	
	DBBVF_LSC (*DB, LB, BS*)	94.79±0.29		100.00±0.00		100.00±0.00	
	DBBF_LSC (*DB, LB, BS*)	91.83±1.40		95.50±0.37		98.79±0.41	
	DBBNF_LSC (*DB, LB, BS*)	99.21±0.48		99.45±0.11		99.96±0.09	
	SHB_LSC (*BS, ShC*)	94.26±9.24		0.00±0.00 *		96.82±4.66	
Water or shadow	DPWASH_LSC (*ShC, Wa*)	100.00±0.00		99.99±0.12		100.00±0.00	
	TWASH_LSC (*DB, LB, ShC, Wa*)	96.85±1.07		99.89±0.14		100.00±0.00	
Snow or cloud or bright bare soil	SN_CL_BBB_LSC (*DB, LB, BS, ShC*)	100.00±0.00		99.91±0.03		100.00±0.00	
	In parentheses are the *Reference Classes* according to Table 10, Producer's Accuracy, $p_{prdcr,c}\pm\delta_c$, c = 1, ..., *RC* = 7: refer to equations (21) and (26), with α = RC/100, (1- α/RC) = 0.99, χ^2 = 6.63.	**QuickBird-2**		**WorldView-2, T1**		**WorldView2, T2**	
		$p_{p,c}$	Pixels and (polygons), if any	$p_{p,c}$	Pixels and (polygons), if any	$p_{p,c}$	Pixels and (polygons), if any
	Dark Built-up (*DB*)	99.87±0.48	79863 (370)	99.85±0.52	103784 (371)	99.30±1.11	98583 (376)
	Light Built-up (*LB*)	99.99±0.13	56107 (372)	99.95±0.30	96608 (370)	99.24±1.16	90522 (371)
	Tree Crown (*TCrwn*)	99.96±0.27	71012 (370)	99.84±0.53	76568 (379)	99.96±0.26	67675 (401)
	Bare Soil (*BS*)	99.70±0.55	656	99.25±1.11	399	98.57±1.02	906
	Grass (Gr)	99.51±0.73	613	100.00±0.00	437	100.00±0.00	598
	Shadow/Cloud shadow/Cloud (*ShC*)	94.49±2.76	454	98.11±1.82	371	99.56±0.65	682
	Water (*Wa*)	99.30±0.90	573	100.00±0.00	442	99.56±0.80	452

TABLE XII
(Continued.) (b) OVERALL, USER'S, AND PRODUCER'S ACCURACY FOR THE SIAM™ INTERMEDIATE SEMANTIC GRANULARITY CLASSIFICATION OF TEST DATA SETS REPRESENTED IN PERCENTAGES. SYMBOL ^—ERROR LIKELY DUE TO EFFECTS OF RELATIVE CALIBRATION PROCEDURE OR SENSOR SATURATION

		QuickBird-2		WorldView-2, T1		WorldView-2, T2	
Q-SIAM™ super-categories at intermediate semantic granularity	**Overall Accuracy, $p_{OA}\pm\delta$: refer to equations (19) and (24), with α = 0.05, (1-α) = 0.95. Inter-vocabulary degree of match *CVPSI* = 0.6134, with *TC* = 28, refer to Table 11.**	98.97 ± 0.34		99.57 ± 0.24		99.44 ± 0.24	
Q-SIAM™ super-categories at intermediate semantic granularity, pseudo-colors	**SIAM™ spectral categories at intermediate granularity, pseudo-colors and acronyms (in parentheses are the *Reference Classes* according to Table 10), User's Accuracy, $p_{usr,c}\pm\delta_c$, c = 1, ..., *TC* = 28: refer to equations (21) and (25), with α = TC/100, 1- α/TC = 0.99, χ^2 = 6.63.**	**QuickBird-2**		**WorldView-2, T1**		**WorldView-2, T2**	
Vegetation	SV_SC (*TCrwn, Gr*)	100.00±0.00		100.00±0.00		99.99±0.03	
	AV_SC (*TCrwn, Gr, ShC*)	100.00±0.00		100.00±0.00		99.79±0.46	
	SHRWE_LSC (*TCrwn, ShC*)	99.65±0.94		100.00±0.00		99.90±0.26	
	SHV_WEDR_LSC (*TCrwn, ShC*)	93.84±3.98		98.70±1.05		50.74±3.00^	
	ASHRBR_SC (*TCrwn, Gr*)	98.07±0.17		99.21±0.09		99.50±0.07	
	AHRBCR_LSC (*TCrwn, Gr*)	0.00±0.00*		0.00±0.00*		100.00±0.00	
	PB_LSC (*TCrwn, Gr*)	97.35±0.81		94.24±4.74		95.47±1.20	
	GH_CL_LSC (*TCrwn, Gr, ShC*)	98.58±13.62*		0.00±0.00*		0.00±0.00*	
Bare soil or built up	BBB_VBBB_SC (*LB, BS*)	99.99±0.06		100.00±0.00		100.00±0.00	
	SADBBVF_SC (*DB, LB, BS*)	90.57±0.30		100.00±0.00		100.00±0.00	
	SADBBF_SC (*DB, LB, BS*)	94.40±0.79		96.68±0.31		99.19±0.27	
	SADBBNF_SC (*DB, LB, BS*)	99.44±0.17		99.97±0.02		99.99±0.02	
	SHB_LSC (*BS, ShC*)	94.26±9.24		0.00±0.00*		96.82±4.66	
Water or shadow	DPWASH_LSC (*ShC, Wa*)	100.00±0.00		99.99±0.12		100.00±0.00	
	TWASH_LSC (*DB, LB, ShC, Wa*)	96.85±1.07		99.89±0.14		100.00±0.00	
Snow or cloud or bright bare soil	SN_CL_BBB_LSC (*DB, LB, BS, ShC*)	100.00±0.00		99.91±0.03		100.00±0.00	
	In parentheses are the *Reference Classes* according to Table 10, Producer's Accuracy, $p_{prdcr,c}\pm\delta_c$, c = 1, ..., *RC* = 7: refer to equations (21) and (26), with α = RC/100, 1- α/RC = 0.99, χ^2 = 6.63.	**QuickBird-2**		**WorldView-2, T1**		**WorldView-2, T2**	
		$p_{prdcr,c}$	Pixels and (polygons), if any	$p_{prdcr,c}$	Pixels and (polygons), if any	$p_{prdcr,c}$	Pixels and (polygons), if any
	Dark Built-up (DB)	99.87±0.48	79863 (370)	99.85±0.52	103784 (371)	99.30±1.11	98583 (376)
	Light Built-up (LB)	99.99±0.13	56107 (372)	99.95±0.30	96608 (370)	99.24±1.16	90522 (371)
	Tree Crown (TCrwn)	99.96±0.27	71012 (370)	99.84±0.53	76568 (379)	99.96±0.26	67675 (401)
	Bare Soil (BS)	99.70±0.55	656	99.25±1.11	399	99.67±0.49	906
	Grass (Gr)	99.51±0.73	613	100.00±0.00	437	100.00±0.00	598
	Shadow/Cloud shadow/Cloud (ShC)	94.49±2.76	454	98.11±1.82	371	99.56±0.68	682
	Water (Wa)	99.30±0.90	573	100.00±0.00	442	99.56±0.80	452

correlation value existing between a possible training data set and the test data set selected in these experiments, because SIAM™ employs no training data set at all (refer to Section II-G). This thematic map accuracy assessment scenario is totally different from that typically faced by RS scientists and practitioners involved with the quality assessment of thematic maps generated from RS imagery by traditional inductive supervised data learning classifiers (statistical classifiers, refer to Section II-B), e.g., artificial neural networks, radial basis functions, support vector machines, non-parametric nearest-neighbor classifiers, parametric maximum likelihood classifiers, etc. [32]–[34]. In these cases, a test data set is a subset of the reference supervised (labeled) data independent of the training data set, but belonging to the same probability distribution. If an inductive data learning classifier performs well with the training data set, then it is good in learning how to guess labels of the training samples. For example, if no incorrect prediction of labels is made during training, then this supervised data learning classifier is termed *consistent classifier* [132], [133]. If the supervised data learning classifier performs with the training data much better than with the test data, then it lacks generalization capability. If the training and test data sets are the same set, it is obvious that the performance of the inductive classifier in the testing phase would be as high as in the training phase, but the classifier "true" generalization capability would remain unknown. Unfortunately, accuracy estimates of classification maps generated by inductive data learning classifiers proposed to the RS community are: 1) rarely provided with a degree of uncertainty in measurement (as a negative example not to be imitated, see [134]), which violates well-known laws

TABLE XII
(Continued.) (c) OVERALL, USER'S, AND PRODUCER'S ACCURACY FOR THE SIAM™ COARSE SEMANTIC GRANULARITY CLASSIFICATION OF TEST DATA SETS REPRESENTED IN PERCENTAGES. SYMBOL ^—ERROR LIKELY DUE TO EFFECTS OF RELATIVE CALIBRATION PROCEDURE OR SENSOR SATURATION

		QuickBird-2		WorldView-2, T1		WorldView-2, T2	
Q-SIAM™ super-categories at coarse semantic granularity	**Overall Accuracy, $p_{OA}\pm\delta$: refer to equations (19) and (24), with α = 0.05, (1-α) = 0.95. Inter-vocabulary degree of match *CVPSI* = 0.5136, with *TC* = 12, refer to Table 11.**	99.32 ± 0.36		99.76 ± 0.24		99.65 ± 0.25	
Q-SIAM™ super-categories at coarse semantic granularity, pseudo-colors	**SIAM™ spectral categories at coarse granularity, pseudo-colors and acronyms (in parentheses are the *Reference Classes* according to Table 10), User's Accuracy, $p_{usr,c}$ $\pm\delta_c$, c = 1, ..., *TC* = 12: refer to equations (21) and (25), with α = TC/100, (1- α/TC) = 0.99, χ^2 = 6.63.**	**QuickBird-2**		**WorldView-2, T1**		**WorldView-2, T2**	
Vegetation	VGT (*TCrwn, Gr, ShC*)	100.00±0.00		100.00±0.00		99.86±0.12	
	SHV (*TCrwn, ShC*)	99.48±0.83		99.81±0.35		99.08±0.46	
	RNGLND (*TCrwn, Gr, ShC*)	98.07±0.17		99.21±0.09		99.50±0.07	
	PB (*TCrwn, Gr*)	97.35±0.81		94.24±4.74		95.47±1.20	
	GH (*TCrwn, Gr, ShC*)	98.58±13.62*		0.00±0.00*		0.00±0.00*	
Bare soil or built up	BB (*DB, LB, BS*)	99.80±0.04		99.98±0.01		99.98±0.01	
	SHB (*BS, ShC*)	94.26±9.24		0.00±0.00*		96.82±4.66	
Water or shadow	WASH (*DB, LB, ShC, Wa*)	99.85±0.21		99.98±0.06		100.00±0.00	
Snow or cloud or bright bare soil	SN_CL_BBB (*DB, LB, BS, ShC*)	100.00±0.00		99.91±0.03		100.00±0.00	
	In parentheses are the *Reference Classes* according to Table 10, Producer's Accuracy, $p_{prdcr,c}$, $\pm\delta_c$, c = 1, ..., *RC* = 7: refer to equations (21) and (26), α = RC/100, (1- α/RC) = 0.99, χ^2 = 6.63.	**QuickBird-2**		**WorldView-2, T1**		**WorldView-2, T2**	
		$p_{prdcr,c}$	Pixels and (polygons), if any	$p_{prdcr,c}$	Pixels and (polygons), if any	$p_{prdcr,c}$	Pixels and (polygons), if any
	Dark Built-up (DB)	99.87±0.48	79863 (370)	99.85±0.52	103784 (371)	99.30±1.11	98583 (376)
	Light Built-up (LB)	99.99±0.13	56107 (372)	99.95±0.30	96608 (370)	99.24±1.16	90522 (371)
	Tree Crown (TCrwn)	99.96±0.27	71012 (370)	99.84±0.53	76568 (379)	99.96±0.26	67675 (401)
	Bare Soil (BS)	99.70±0.55	656	99.25±1.11	399	99.67±0.49	906
	Grass (Gr)	99.51±0.73	613	100.00±0.00	437	100.00±0.00	598
	Shadow/Cloud shadow/Cloud (ShC)	96.92±2.09	454	99.46±0.98	371	99.85±0.38	682
	Water (Wa)	99.30±0.90	573	100.00±0.00	442	99.56±0.80	452

of sample statistics [50], [51], together with common sense envisaged by the QA4EO guidelines [3], 2) extracted from one or two RS images at most where both training and test data sets are identified, which is equivalent to a toy problem unable to stress the algorithm's robustness to changes in the input data set, and 3) rarely provided with any proof of the required independence between the training and testing data sets. To avoid these experimental drawbacks, this paper investigates the thematic accuracy of the deductive Q-SIAM™ preliminary classification maps automatically generated from VHR images by means of: 1) a set of mutually uncorrelated TQIs provided with an error tolerance in compliance with the QA4EO guidelines and 2) three VHR test images acquired by two different sensors where no training of the Q-SIAM™ classifier ever occurs, therefore statistical independence between the training and testing data sets is guaranteed (because no training data set exists).

To recapitulate, the TQI values reported in Tables XII(a)–(c) score "high": 1) without any bias due to a possible correlation between training and testing data sets because, SIAM™, which is physical model based (prior knowledge based), employs no training data at all, 2) in compliance with theoretical expectations about the accuracy of a thematic map in relation to its degree of semantic specialization (or, *vice versa*, generalization), 3) in agreement with the quantitative CVPSI values estimated, in compliance with the QA4EO guidelines, as a source of numerical evidence independent of TQIs, and 4) in line with the high-value OQIs (refer to Section II-D) claimed for SIAM™ by the existing literature [5]–[17].

A final experimental remark regards the applicability of an automatic real-time SIAM™-based bi-temporal post-classification change detection approach to VHR imagery, see Section IV-C. The OA of a bi-temporal post-classification change/no-change detection map is upper bounded by (8), which makes a thematic map pair difference recommended if and only if the two categorical maps employed as input are very accurate. Given the two overall accuracies collected from the test WV-2 T1 and T2 image pair at fine semantic granularity, (8) is instantiated as follows:

$$\begin{aligned}&\text{Accuracy of the automatic real-time SIAM}^{\text{TM}}\text{-based bi-temporal post-classification change/no-change detection map}\\ &<= (\text{Accuracy of the map at time T1} \times \text{Accuracy of the map at time T2})\\ &= (99.57\% \pm 0.24\%) \times (99.45\% \pm 0.24\%)\\ &= [99.33\%, 99.81\%] \times [99.20\%, 99.68\%]\\ &= [98.53\%, 99.49\%].\end{aligned}$$

This makes the near real-time post-classification change detection based on SIAM™ maps theoretically feasible.

Unfortunately, no quantitative assessment of the change/no-change maps shown in Fig. 11(a) and (b) is carried out in this experimental work, due to the absence of any ground-truth information collected from a reference data source one step closer to the ground than the VHR images used to make up the change maps (refer to Section III).

2) Spatial Accuracy Assessment: Accuracy assessment of thematic maps generated from VHR imagery requires, in addition to the estimation of pixel-based TQIs, the estimation of polygon-specific SQIs (refer to Section III). In other words, SQI assessment must be driven by sampling units whose spatial type is polygon (refer to Section VI-D1).

In general, a data set of reference image-objects can employ polygons as sampling units according to the following criteria [60]:

- Reference image-objects have well-defined geometric properties (e.g., buildings typically feature high rectangularity), thus they can be modeled as polygons fitting their shape.
- Reference image-objects can be delineated with a reasonable level of effort.
- The probability that a reference image-object is selected is known (see Section V). The area of each polygon varies on a per-object basis; therefore, the use of polygons as reference samples merits special consideration [123]. Variability in object area influences the probability of selecting a given object with respect to other objects of the same reference class. To compensate for these varying inclusion probabilities, it is necessary to adjust the inclusion probability for each object by an inverse area function, see (23) in Section VI-B4 [60].
- Inclusion probabilities for non-sampled reference image-objects are knowable (see Section V).

For example, Table IX lists the reference LC classes whose sampling units are polygons according to the aforementioned criteria. Per-class SQIs collected over class-specific reference objects, say, reference objects belonging to either class DB, LB or $TCrwn$ (refer to Table IX), must be computed as a weighted sum, where each reference object-specific weight is provided by (23).

In this paper, SQIs adopted for the spatial quality assessment of sub-symbolic (non-semantic) segmentation maps univocally generated from (symbolic) classification maps (refer to footnote 1 and also refer to Sections II-G and III) are adapted from the existing literature, e.g., refer to [54], [101], [102]. The proposed set of SQIs consists of:

- an oversegmentation QI (OSQI), where $OSQI$ values belong to range [0, 1],
- an undersegmentation QI (USQI), where $USQI$ values belong to range [0, 1], and
- a pair of fuzzy edge overlap QIs (FEOQIs), where $FEOQI$ values belong to range [0, 1].

For each reference class $c = 1, \ldots, RC$ (see Table IX), reference object-specific spatial accuracy values $OSQI_{i,c}$, $USQI_{i,c}$, and $FEOQI_{i,c}$ are measures of the relationship between the i-th reference image-object (polygon) belonging to class c, identified as $RO_{i,c}$, and its corresponding *mapped image-object*, also called test image-object, identified as $TO_{i,c}$. For a given reference object, $RO_{i,c}$, the single 2-D object in the test map to be selected for comparison purposes, $TO_{i,c}$, is the one providing the most pixels in common with the reference object $RO_{i,c}$[54]. Therefore

$$TO_{i,c} = \underset{\forall TO_j \in TO}{Arg} \max |RO_{i,c} \cap TO_j|, c = 1, \ldots, RC. \quad (27)$$

The reference object-specific value $OSQI_{i,c}$ quantifies the area of overlap of the object pair, $(RO_{i,c}, TO_{i,c})$, with respect to the reference object $RO_{i,c}$, where expressions $(RO_{i,c} \cap TO_{i,c}) \subseteq RO_{i,c}$ and $(|RO_{i,c} \cap TO_{i,c}|/|RO_{i,c}|) \leq 1$ hold and where operator $|\cdot|$ computes a set's cardinality, thus:

$$OSQI_{i,c}(RO_{i,c}, TO_{i,c}) = \frac{|RO_{i,c} \cap TO_{i,c}|}{|RO_{i,c}|} \in [0, 1], \quad c = 1, \ldots, RC. \quad (28)$$

The $OSQI_{i,c}(RO_{i,c}, TO_{i,c})$ scalar value in range [0, 1] must be maximized (up to 1) (in [54], a dual entity of (28), namely, an oversegmentation error estimate, must be minimized to 0).

Similarly, the reference object-specific value $USQI_{i,c}$ quantifies the area of overlap of the object pair, $(RO_{i,c}, TO_{i,c})$, with respect to the test object $TO_{i,c}$, where expressions $(RO_{i,c} \cap TO_{i,c}) \subseteq TO_{i,c}$ and $(|RO_{i,c} \cap TO_{i,c}|/|TO_{i,c}|) \leq 1$ hold, thus:

$$USQI_{i,c}(RO_{i,c}, TO_{i,c}) = \frac{|RO_{i,c} \cap TO_{i,c}|}{|TO_{i,c}|} \in [0, 1], \quad c = 1, \ldots, RC. \quad (29)$$

The $USQI_{i,c}(RO_{i,c}, TO_{i,c})$ scalar value in range [0, 1] must be maximized (up to 1) (in [54], a a dual entity of (29), namely, an undersegmentation error estimate, must be minimized to 0).

The FEOQI-Reference (FEOQI-R) index measures the precision of the reference object edges recognized in the test map with respect to the total number of edge pixels in the reference object. Its dual SQI, called FEOQI-Test (FEOQI-T) index, measures the precision of the reference object edges recognized in the test map with respect to the total number of edge pixels in the test object. To compensate for mixed pixel effects due to the interaction between surface phenomena and the spatial resolution of the imaging sensor, together with human errors in identifying image-objects and localizing their boundaries, a tolerance in the recognition of the image borders is introduced as an appropriate buffer zone distance, d. In practice, parameter d is the width of the extracted border line of both the test and the reference object. FEOQI-R has the appealing quality of assessing the overlapping area between the mapped object and the reference object with regard to the reference object while remaining somewhat robust to the adjacency (neighboring) issues which deeply affect the USQI and OSQI index estimations (see below in this section). Unfortunately, unlike FEOQI-R, FEOQI-T is affected by the same adjacency issues affecting the USQI and OSQI indexes. Let $e(\cdot)$ denote the operator that extracts the set of edge pixels from a generic reference region

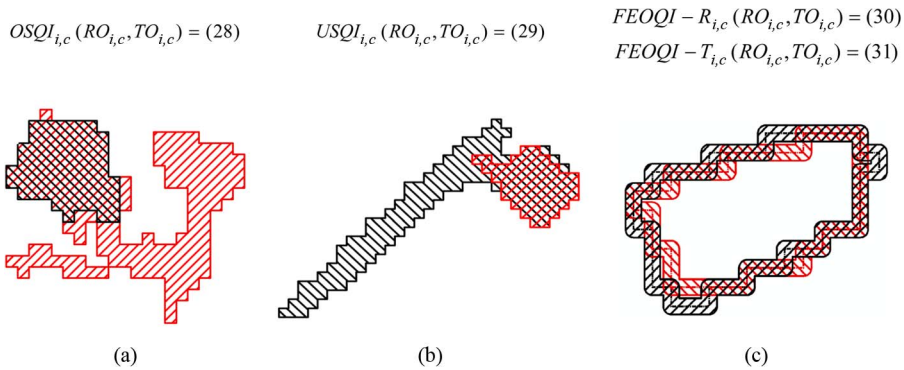


Fig. 13. Illustration of SQIs. Reference object, RO_i, shown in red, and mapped (test) object, TO_i, shown in black. (a) Low $OSQI_i(RO_i, TO_i)$ example. (b) Low $USQI_i(RO_i, TO_i)$ example. (c) Example where $1 \geq FEOQI - R_{i,c}(RO_{i,c}, TO_{i,c}) > FEOQI - T_{i,c}(RO_{i,c}, TO_{i,c}) \geq 0$.

RO_i and test region TO_i. Thus [54]:

$$FEOQI - R_{i,c}(RO_{i,c}, TO_{i,c}) = \frac{|e(RO_{i,c}) \cap e(TO_{i,c})|}{|e(RO_{i,c})|} \in [0,1], c = 1, \ldots, RC. \quad (30)$$

The $FEOQI - R_{i,c}(RO_{i,c}, TO_{i,c})$ scalar value in range [0, 1] must be maximized (up to 1) (in [54], a dual entity of (30), namely, an edge location error, must be minimized to 0)

$$FEOQI - T_{i,c}(RO_{i,c}, TO_{i,c}) = \frac{|e(RO_{i,c}) \cap e(TO_{i,c})|}{|e(TO_{i,c})|} \in [0,1], c = 1, \ldots, RC. \quad (31)$$

The $FEOQI - R_{i,c}(RO_{i,c}, TO_{i,c})$ scalar value in range [0, 1] must be maximized (up to 1). It is noteworthy that no parameter equivalent to $FEOQI - T_{i,c}$ is proposed in [54].

Examples of image-objects involved with the computation of SQIs via (28)–(31) are shown in Fig. 13.

To conclude, *it is important to stress that, whereas pixel-based TQIs estimated in this work are based on a realistic many-to-many association model between the test and reference semantic vocabularies in accordance with the hypothesis of dealing with the second type of maps comparison described in* Section VI-A, *polygon-specific SQIs are estimated according to* (27) *up to* (31), *where a one-to-one association model between one reference and one mapped polygon holds, see* (27). *This latter constraint is unable to capture any possible "correct" association of one polygon belonging to a specific reference class with one or more polygons belonging to a set of "correct" test classes*. This means that, based on theoretical considerations exclusively, in the second type of inter-map comparisons described in Section VI-A, where many-to-many semantic relations must be considered "*correct*," SQIs computed according to (27)–(31) are expected to be negatively biased (i.e., underestimated), whereas the same SQI formulas have no bias when applied to the first type of inter-map comparisons described in Section VI-A. For example, if one reference LC class is associated with several test categories, like in Table X, then the estimated TQI values of this reference class are expected to be "higher" (due to their inherent semantic vagueness, refer to Section VI-D1) than its SQIs, computed through (27)–(31), which are negatively biased due to the assumption, adopted by (27), that the "*correct*" relation between a reference object and a mapped object belonging to the test map domain is one-to-one.

Instantiation with the SIAM™ pre-classification maps automatically generated from the WV-2 and QB-2 test images: Multi-scale image segmentation maps can be automatically generated in near real-time from the SIAM™ preliminary classification maps featuring multiple semantic granularities (see Table V) by means of a traditional well-posed two-pass connected-component image labeling algorithm [20] (refer to footnote 1 in Section I). This is an improvement over traditional multi-scale image segmentation algorithms, like that proposed by Baatz *et al.* [38] and implemented at the first stage of two-stage non-iterative GEOBIA systems and three-stage iterative GEOOIA systems, like the popular Definiens Developer commercial software product [89], [90] (refer to Section II-F). First, SIAM™ delivers as output in near real-time semantic (symbolic) multi-scale object-based parent–child relations whereas traditional multi-scale image segmentation algorithms generate as output sub-symbolic (non-semantic) multi-scale image-objects exclusively. Thus, the symbolic information conveyed by the former is potentially richer than the sub-symbolic information provided by the latter (refer to Section II-A). Second, whereas SIAM™ is automatic, requiring no user-defined parameter to run (refer to Section II-G), the latter require free parameters to be user-defined based on heuristics. For example, the iterative multi-scale image segmentation algorithm proposed by Baatz *et al.* [38], [89], [90] employs at least three (actually more, e.g., see [117]) parameters to be user-defined based on empirical criteria, namely, a scale parameter < 0 (such that increasing scale parameter values will result in the detection of a smaller number of larger image-objects; in practice, this so-called spatial scale parameter is an upper bound on the spectral variance of image-objects, whose physical meaning and unit of measure have nothing to do with a spatial scale), shape versus color weight $\in$ [0, 1] and shape compactness versus shape smoothness $\in$ [0, 1] [38], [142].

To summarize, the capability of SIAM™ to automatically generate as output in near real-time multi-scale image segmentation maps in parallel with multi-granularity preliminary classification maps makes SIAM™ a viable alternative to traditional semi-automatic, multi-scale, sub-symbolic image segmentation algorithms implemented at the first stage of state-of-the-art GEOBIA/ GEOOIA systems [38], [89]–[92], [131], [142], refer to Section II-F.

The spatial quality assessment of the Q-SIAM™ maps, generated from the WV-2 and QB-2 test images described in Section IV, is instantiated according to the following constraints.

(a) A reference class-specific sample unit, $s_c \in S_c \subseteq S \equiv$ GEOROI is a sample polygon (s_plygn) for class $c = DB, LB, TCrwn$, see Table IX, selected according to the non-standard class-specific SIRS strategy shown in Fig. 12 (refer to Section VI-B4).

(b) A class-specific sample set size SSS_c is estimated as $SSS_c =$ (22) (refer to Section VI-B1), where samples are s_plygns $s_{i,c}$, $i = 1, \ldots, SSS_c$, $c = DB, LB, TCrwn$, see Table IX.

(c) The four reference polygon-specific SQIs defined by (27)–(31), must be computed for each of the three reference classes whose sampling unit type is polygon, namely, classes $c = DB, LB, TCrwn$, see Table IX, and for each class-specific s_plygn with index $i = 1, \ldots, SSS_c$; in other words, class- and polygon-specific quality index values $OSQI_{i,c}, USQI_{i,c}$, $FEOQI - R_{i,c}$ and $FEOQI - T_{i,c}$ must be computed with $c = DB, LB, TCrwn$, and $i = 1, \ldots, SSS_c$, according to (27)–(31).

(d) The target population U consists of pixels $u \in U \equiv S \equiv$ GEOROI, where GEOROI comprises the whole test map made up from a VHR image.

(e) To account for the unequal inclusion probability of s_plygns in compliance with the Horvitz–Thompson theorem, such that (13) tends to coincide with (15), i.e., (13)→(15) holds (see Section V), then for each pixel $u \in U$ such that $u \cap s_{i,c} \neq 0$, $c \in \{DB, LB, TCrwn\}$, $i \in \{1, SSS_c\}$, the u's target population pixel-specific inclusion probability $\pi_{u,i,c}$ must be $\pi_{u,i,c} =$ (23) so that the u's pixel-specific weight $w_{u,i,c}$ becomes $w_{u,i,c} =$ (14) $= 1/$(23), i.e., *all pixels belonging to the same s_plygn share the same inclusion probability which is polygon-specific according to* (23).

The aforementioned constraints (a) to (e) mean that the four reference class-specific image-wide pixel-based SQIs, namely, $OSQI_c$, $USQI_c$, $FEOQI - R_c$, and $FEOQI - T_c$, $c = DB, LB, TCrwn$ (see Table IX), must be estimated as the weighted sum of the class- and polygon-specific index values $OSQI_{i,c} =$ (28), $USQI_{i,c} =$ (29), $FEOQI - R_{i,c} =$ (30), and $FEOQI - T_{i,c} =$ (31), $i = 1, \ldots, SSS_c =$ (22), $c = DB, LB, TCrwn$, computed across pixels belonging to the class-specific s_plygns $RO_{i,c}$, $i = 1, \ldots, SSS_c$, $c = DB, LB, TCrwn$, where the pixel-specific weight is computed as the inverse of (23) (refer to Section VI-B1) in compliance with the Horvitz–Thompson theorem, see (14) (refer to Section V). For example

$$\begin{aligned}
&\text{Pixel-based } OSQI_c \in [0,1], c = DB, LB, TCrwn \\
&= \frac{1}{\left(\sum_{j=1}^{SSSc} w_{j,c} \cdot area_{j,c}\right)} \cdot \sum_{i=1}^{SSSc} w_{i,c} \cdot area_{i,c} \\
&\quad \cdot \text{Polygon-specific } OSQI_{i,c}(RO_{i,c}, TO_{i,c}), \\
&c = DB, LB, TCrwn, \\
&= \frac{1}{\left(\sum_{j=1}^{SSSc} \frac{1}{\pi_{j,c}} \cdot area_{j,c}\right)} \sum_{i=1}^{SSSc} \frac{1}{\pi_{i,c}} \cdot area_{i,c} \cdot \\
&\quad OSQI_{i,c}(RO_{i,c}, TO_{i,c}), c = DB, LB, TCrwn, \\
&= \frac{1}{\left(\sum_{h=1}^{SSSc} area_{h,c}\right) \cdot \left(\sum_{j=1}^{SSSc} \frac{1}{area_{j,c}} \cdot area_{j,c}\right)} \\
&\quad \times \sum_{i=1}^{SSSc} \frac{\left(\sum_{h=1}^{SSSc} area_{h,c}\right)}{area_{i,c}} \cdot area_{i,c} OSQI_{i,c}(RO_{i,c}, TO_{i,c}), \\
&c = DB, LB, TCrwn, \\
&= \frac{1}{SSS_c} \sum_{i=1}^{SSSc} OSQI_{i,c}(RO_{i,c}, TO_{i,c}), \\
&c = DB, LB, TCrwn
\end{aligned} \tag{32}$$

where

Polygon-specific $OSQI_{i,c}(RO_{i,c}, TO_{i,c})$ $= (28) \in [0,1], \forall i \in \{1, SSS_c\}$, and pixel-based $OSQI_{i,c}$ $\in [0,1], c = DB, LB, TCrwn$.

Equation (32) *shows that, when the pixel-specific weight is computed according to* (23), *then a class-specific image-wide pixel-based* SQI_c, $c = 1, \ldots, RC$, *is equivalent to the mean of the SQIs computed per reference and mapped 2-D object pair*, $SQI_{i,c}(RO_{i,c}, TO_{i,c})$, $i = 1, \ldots, SSS_c$, $c = 1, \ldots, RC$.

To better understand the degree of novelty of the proposed protocol for SQI estimation in thematic maps generated from VHR images, let us compare it with the so-called protocol for accuracy assessment in classification of VHR images proposed in [54]. Irrespective of the fact that SQIs proposed in (27)–(31) are strictly related to a selection of the geometric error indices proposed in [54], no comparison between results collected by these two protocols is possible due to their methodological and implementation differences. These differences are summarized below.

- Whereas the present work employs (23) to compensate for unequal probability sampling, no Horvitz–Thompson theorem is accounted for in [54].
- The protocol proposed in [54] is applied to LC maps generated by supervised data learning classifiers where the reference and test semantic taxonomies are the same (with a cardinality equal to eight). Hence, the protocol for map assessment proposed in [54] refers to the first type of inter-map comparisons described in Section VI-A. The protocol proposed in the present work is more general than that discussed in [54] because the former applies to test and reference spectral vocabularies which may or may not coincide, i.e., it refers to the second type of map pair comparisons described in Section VI-A.
- SQIs computed according to (27)–(31) may be biased due to undesired eight-adjacency neighboring effects. For example, Fig. 14 shows a reference image-object belonging to class LB (see Table IX) affected by an undersegmentation error because its mapped object (test object) spans the entire image as an undesired effect

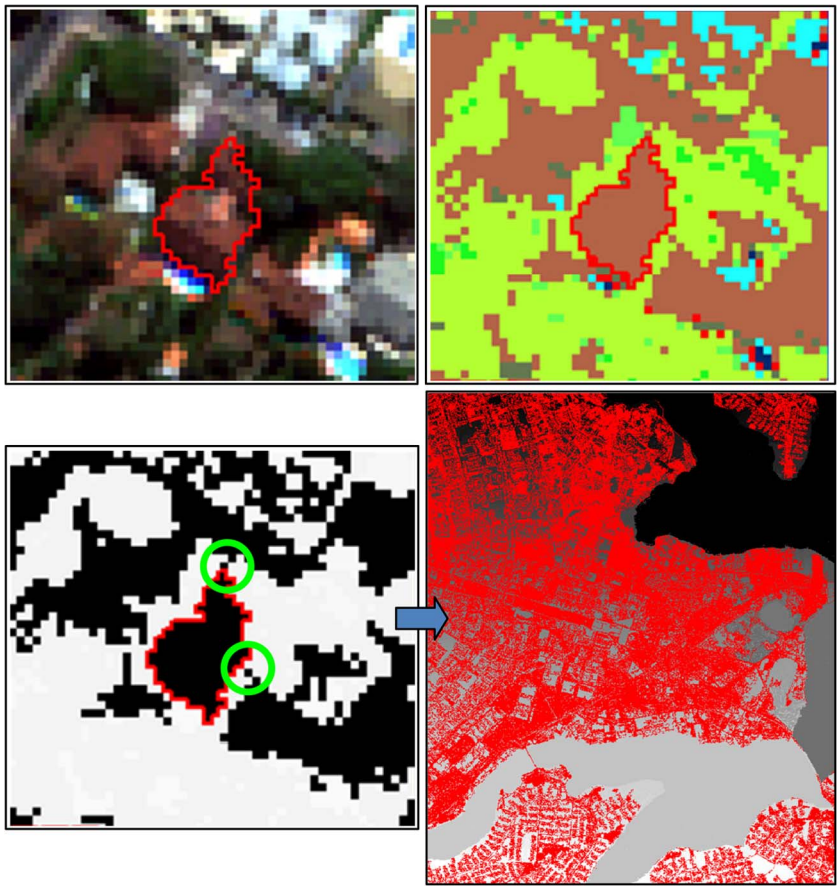


Fig. 14. Reference image-object identification in a SIAM™ coarse granularity map. Top Left. Bare Soil (BS) image-object identified by an expert photointerpreter. Top Right. Reference polygon perimeter overlaid on a SIAM™ coarse map classification. Bottom. Eight-connectivity adjacency (green circles, bottom left) resulting in a single segment (in red, bottom right), spanning the entire image. As a consequence, SQIs of the depicted reference object are: low $USQI$, high $OSQI$ (close to 1), high $FEOQI - R$, and low $FEOQI - T$ values.

of an eight-adjacency neighborhood. In practice, due to this undesired eight-adjacency neighborhood effect, oversegmentation quality, see (28), is overestimated while undersegmentation quality, see (29), and edge position quality, see (30) and (31), are underestimated. The same considerations should hold for the protocol proposed in [54], whose SQIs are highly correlated with those adopted in this work. However, in [54], the aforementioned spatial error overestimation was not observed. This is due to the fact that, in [54], a toy problem, consisting of a small-size segmentation test map and a small-cardinality reference object set (comprising 11 samples) was considered for validation. On the contrary, the present work adopts for protocol validation hundreds of reference samples to be mapped onto real-world segmentation maps generated from test thematic maps of full-size VHR images.

- In [54], the spatial resolution of the test image is 0.7 m, while in the present work the spatial resolution of the test images ranges from the 2.0 m resolution of the WV-2 sensor to the 2.4 m of QB-2 imagery. This implies that the VHR image data set employed in this work is more susceptible to undersegmentation errors when an eight-adjacency neighborhood model is adopted.

To recapitulate, based on both theoretical speculations and practical considerations (see Fig. 14), *the proposed protocol* applied to the assessment of the Q-SIAM™ maps generated from the WV-2 and QB-2 test images described in Section IV *is expected to deliver SQIs which, in general, are lower than TQIs due to a summation of two effects*.

(a) An inadequacy of (27) to cope with a test and a reference semantic vocabulary when they do not coincide. This causes all SQIs computed via (27)–(31) to be underestimated.

(b) An undesired eight-adjacency neighborhood phenomenon (see Fig. 14) which causes: 1) the oversegmentation quality index, see (28), to be overestimated, and 2) the undersegmentation quality index, see (29), together with the edge position quality indexes, see (30) and (31), to be underestimated.

Table XIII shows the SQIs collected for the Q-SIAM™ classification maps generated in Section IV. In line with the aforementioned theoretical and practical considerations, SQIs reported in Table XIII are generally lower than TQIs shown in Tables XII(a)–(c). In greater detail, Table XIII reveals that the reference class LB outperforms the DB and $TCrwn$ classes (refer to Table V) in all test images and at all granularities. Furthermore, the LB data sets provide the most consistent results, with each SQI varying by less than one percent across semantic granularities in each test image. DB areas are highly susceptible to undersegmentation errors, particularly at coarse

TABLE XIII
SQIs FOR FINE, INTERMEDIATE, AND COARSE Q-SIAM™ CLASSIFICATION MAPS OF THE THREE VHR TEST IMAGES, IN PERCENT VALUES. SYMBOL *—BUFFER ZONE DISTANCE PARAMETER, d, IS SET TO 2 PIXELS IN THE FEOQI CALCULATIONS, REFER TO (30) AND (31)

Test dataset		Semantic granularity (Fine: F = 52, Intermediate: I = 28, Coarse: C = 12)	Number of reference objects (polygons)	$OSQI_c \times 100$, see (28) $\pm\delta$ = (18), with 1- α = 0.95.	$USQI_c \times 100$, see (29) $\pm\delta$ = (18), with 1- α = 0.95.	$FEOQI\text{-}R_c^* \times 100$, see (30) $\pm\delta$ = (18), with 1- α = 0.95.	$FEOQI\text{-}T_c^* \times 100$, see (31) $\pm\delta$ = (18), with 1- α = 0.95.	Percent Average SQI
QB-2	Dark Built-up (*DB*)	C	291	97.71±1.71	20.98±4.67	70.60±5.23	17.64±4.37	51.73%
		I	291	75.17±4.96	63.63±5.52	76.53±4.86	57.62±5.67	68.24%
		F	291	56.40±5.69	89.11±3.57	69.02±5.31	78.72±4.70	73.31%
	Light Built-up (*LB*)	C	342	97.22±1.74	89.19±3.29	92.71±2.75	88.12±3.42	91.81%
		I	342	97.22±1.74	89.19±3.29	92.71±2.75	88.12±3.42	91.81%
		F	342	97.22±1.74	89.19±3.29	92.71±2.75	88.12±3.42	91.81%
	Tree Crown (*TCrwn*)	C	345	71.19±4.77	50.68±5.27	69.47±4.85	43.71±5.23	58.76%
		I	345	68.29±4.91	54.58±5.25	67.53±4.94	46.69±5.26	59.27%
		F	345	33.28±4.97	82.92±3.97	51.49±5.27	65.82±5.00	58.38%
WV-2, T1	Dark Built-up (*DB*)	C	250	97.16±2.05	16.27±4.57	64.86±5.91	13.09±4.18	47.85%
		I	250	68.94±5.73	61.47±6.03	76.72±5.23	52.38±6.19	64.88%
		F	250	52.07±6.19	88.66±3.93	66.61±5.84	75.58±5.32	70.73%
	Light Built-up (*LB*)	C	352	84.13±3.81	93.78±2.52	86.12±3.61	85.44±3.68	87.37%
		I	352	84.10±3.82	94.02±2.47	86.18±3.60	85.65±3.66	87.49%
		F	352	84.06±3.82	94.22±2.43	86.13±3.61	85.77±3.64	87.55%
	Tree Crown (*TCrwn*)	C	370	88.93±3.19	21.22±4.16	60.20±4.98	18.72±3.97	47.27%
		I	370	88.89±3.20	21.49±4.18	60.28±4.98	18.90±3.98	47.39%
		F	370	50.87±5.09	64.75±4.86	68.90±4.71	54.72±5.07	59.81%
WV-2, T2	Dark Built-up (*DB*)	C	253	98.43±1.53	5.64±2.84	63.84±5.92	4.93±2.66	43.21%
		I	253	80.30±4.90	49.96±6.16	78.96±5.02	43.43±6.10	63.16%
		F	253	64.29±5.90	81.97±4.73	75.27±5.31	70.85±5.59	73.10%
	Light Built-up (*LB*)	C	341	88.48±3.38	92.92±2.72	89.93±3.19	88.89±3.33	90.06%
		I	341	88.42±3.39	93.22±2.66	89.84±3.20	89.18±3.29	90.17%
		F	341	88.38±3.40	93.47±2.62	89.67±3.23	89.30±3.28	90.21%
	Tree Crown (*TCrwn*)	C	380	86.64±3.42	75.76±4.30	88.42±3.21	71.38±4.54	80.55%
		I	380	86.57±3.42	75.72±4.31	88.43±3.21	71.38±4.54	80.53%
		F	380	34.11±4.76	92.48±2.65	54.28±5.00	77.95±4.16	64.71%

semantic granularities, while $TCrwn$ is most susceptible to oversegmentation and edge errors. Table XIII also shows that $OSQI$ and $FEOQI - R$ index values decrease with semantic granularity while $USQI$, $FEOQI - T$, and average SQI increase with the Q-SIAM™ semantic granularity. Oversegmentation and edge errors with respect to the reference objects occur as a result of multiple pixel classifications occurring within the same reference object, see Fig. 15. For example, an object receiving direct sunlight on one side and no sunlight on the other may be assigned with two different spectral category indices although they occur in the same reference thematic object. This issue is less apparent in the coarser levels of granularity because semantic aggregation of mapped classes provides a level of robustness to minor semantic discrepancies (e.g., [SVVH2NIR_LSC ("*Strong Vegetation with Very High 2 Near-Infrared Leaf Spectral Category*"), SVVH1NIR_LSC, SVVH-NIR_LSC] $\subseteq$ [SV_SC ("*Strong Vegetation (Parent) Spectral Category*")] $\subseteq$ [VGT ("*Vegetation Super Spectral Category*")] for 52, 28, and 12 semantic granularities, respectively).

Undersegmentation error is often a result of eight-connectivity neighboring effects, causing pixels connected by one corner to be assigned to the same segment. This error is most common in the reference class DB for the coarse granularity maps, where multiple spatially disjoined reference objects may belong to the same segment, see Fig. 14. As an example, $USQI$ and $FEOQI - R$ index values for the Q-SIAM™ coarse semantic granularity map generated for the WV-2 T2 image are very low. This is due to the eight-connectivity adjacency property described above, resulting in 220 of 253 reference objects being assigned to the same target object (segment 1). This phenomenon holds true for the other test images as well, though it is particularly noticeable in the WV-2 T2 image due to the increased presence of bare soil during the dry season.

The spatial accuracy of the reference class $TCrwn$ for the coarse and intermediate maps changes significantly with the phenological season. SQI values of the reference class $TCrwn$ increase greatly for the WV-2 T2 image due to greater spectral difference between foreground (tree crown) and background (bare soil) during the dry season than exists between the foreground (tree crown) and background (grass) in the green season depicted in the test image WV-2 T1.

To be further investigated in future works, a possible improvement to the proposed protocol for estimation of SQIs, which are biased due to the summation of the two aforementioned effects, may come with a change in the mapped object selection criterion, see (27), required to be capable of modeling many-to-many associations between reference and test classes.

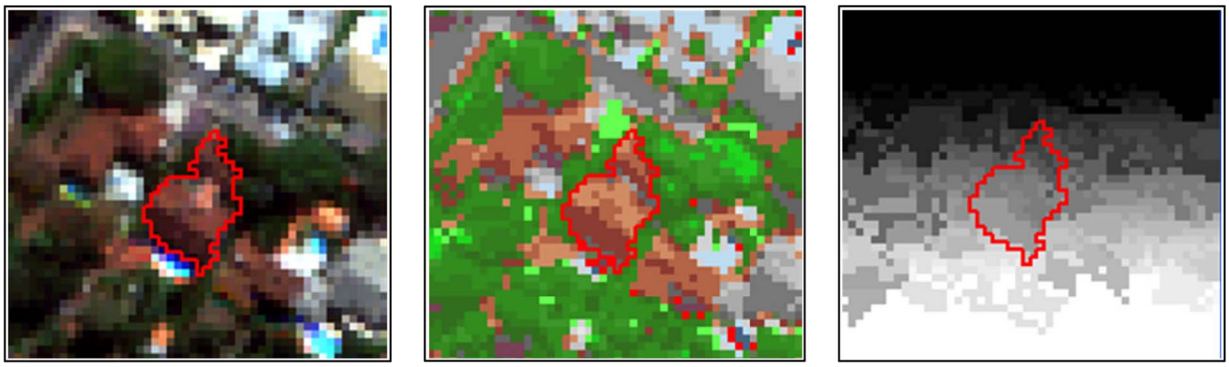


Fig. 15. Reference image-object overlapped with a SIAM™ fine granularity map. Left. Light Built-up (LB) image-object identified by an expert photointerpreter, same as object in Fig. 14. Middle. Reference object overlaid on a SIAM™ fine map classification. Right. Presence of multiple semi-concept labels (numerical identifiers) within the reference object contour results in multiple segments within the object. Note: all labels are in fact plausible spectral descriptions for the reference object. As a consequence, SQIs of the depicted reference object are: high $USQI$, low $OSQI$, low $FEOQI - R$, and low $FEOQI - T$.

VII. Quality Indicators of Operativeness of SIAM™ in the Conducted Experiments

In compliance with the QA4EO guidelines [3], to be considered operational (good-to-go, ready-to-go, turnkey) [26], [135], a satellite-based measurement system must score high in every QI of a set of community-agreed OQIs, such as those listed in Section II-D [5]–[17], [136]. Based on experimental results presented in Section VI, OQIs listed in Section II-D can be instantiated for the (Q-)SIAM™ preliminary classifier as follows.

(a) Degree of automation. SIAM™ is fully automated, i.e., it requires neither user-defined parameters nor reference samples to run. Thus, its ease of use cannot be surpassed.

(b) Effectiveness: classification map accuracy provided with an error tolerance. About the (Q-)SIAM™ TQIs and SQIs provided with a degree of uncertainty in measurement, refer to Section VI in addition to the existing literature [5]–[17]. These TQIs and SQIs must be considered in combination with CVPSI values. The latter provide a measure of the degree of semantic information conveyed by the SIAM™ preliminary classification maps, independent of TQIs and SQIs.

(c) Efficiency: computation time and memory occupation. For example, to map the test images shown in Section IV when running on a Dell laptop featuring an Intel i7 M620 @ 2.67 GHz processor with 8 GB of RAM and a 64-bit Windows 7 operating system, the SIAM™ implementation in the C programming language performs as follows.
- Memory occupation. In these experiments, the SIAM™ dynamic memory size parameter was set equal to 800 MB of RAM.
- Computation time.
- Q-SIAM™, four-band WV-2 image(s), see Section IV, approximately 5000 × 4000 pixels in size. SIAM™ required less than 2 min to generate its complete set of per-image output products including three preliminary classification maps at three levels (fine, intermediate and coarse) of semantic granularity, see Table V.
- Q-SIAM™, four-band QB-2 image, see Section IV, approximately 6400 × 6400 pixels in size. SIAM™ required approximately 2 min to generate its complete set of output products including three preliminary classification maps at fine, intermediate, and coarse semantic granularity, see Table V.
- L-SIAM™, seven-band Landsat-7 ETM+ image(s), see Section IV, 7000 × 8000 pixels in size. SIAM™ required less than 3 min to generate its complete set of per-image output products including three preliminary classification maps at fine, intermediate, and coarse semantic granularity, see Table V.

The parallel implementation of SIAM™ reduces computation time by 15% to 40%, depending on the image size, in a laptop computer with a Windows operating system [16], [17]. An original well-posed two-pass connected-component image labeling algorithm, whose computation time is approximately the same of SIAM™ and increases with the image size, automatically generates multi-scale image segmentation maps from the SIAM™ preliminary classification maps at multiple semantic granularities. Since the time interval between two consecutive spaceborne image acquisitions is not less than approximately 15 min (for the geostationary Meteosat Second Generation satellite), then the available sequential implementation of SIAM™, including its multi-scale image segmentation sub-system, can be considered near real-time.

(d) Robustness to changes in the input data set acquired across time, space, and sensors. In combination with existing literature [5]–[17], TQIs and SQIs collected in Section VI from three VHR test images acquired by two different sensors in two different phenological seasons confirm that SIAM™ appears eligible for use with RS imagery acquired by any existing MS spaceborne/airborne mission provided with radiometric calibration metadata files.

(e) Robustness to changes in input parameters, if any. SIAM™ requires no user-defined parameter, thus its robustness to changes in input parameters cannot be surpassed.

(f) Maintainability/scalability/re-usability to keep up with changes in users' needs and sensor properties. The well-known scalability of SIAM™ to deal with RS imagery acquired by all existing MS spaceborne missions (see Table V) has been confirmed in this experimental work, see Section IV.

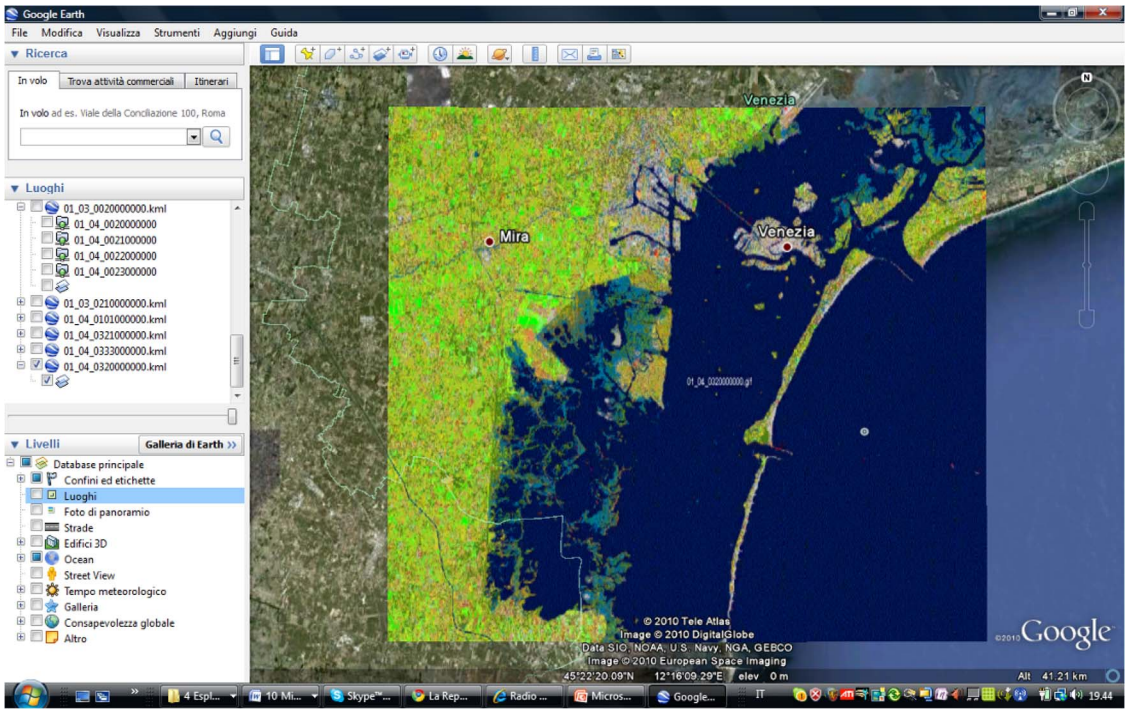


Fig. 16. Preliminary classification map, depicted in pseudo colors, generated by L-SIAM™ from a Landsat-7 ETM+ image of the Venice lagoon, Italy, radiometrically calibrated into TOARF values, spatial resolution: 30 m. The L-SIAM™ map is transformed into the kml data format and loaded as a thematic layer in a commercial 3-D earth viewer (e.g., Google Earth).

(g) Timeliness, defined as the time span between data acquisition and product delivery to the end user. It increases monotonically with manpower, e.g., the manpower required to collect site-specific training samples. Email communications between the first author of this work and DigitalGlobe prove that the following list of activities was performed within one and a half day from the WV-2 data acquisition (while the first two authors were facing other obligations too): WV-2 image calibration, mosaicking, layered stacking and Q-SIAM™ mapping, Landsat image selection, radiometric calibration and L-SIAM™ mapping, WV-2 image re-calibration and Q-SIAM™ re-mapping, QB-2 image calibration and Q-SIAM™ mapping, and, finally, Q-SIAM™-based bi-temporal change detection, refer to Figs. 4(a)–11(b).

(h) Economy (costs, monotonically increasing with manpower and computer power). For example, inductive supervised data learning systems (e.g., nearest-neighbor classifiers, support vector machines, etc. [32]–[34]) increase costs by requiring the collection of reference (training and testing) samples from ground survey, existing maps, ancillary information, etc. Since it is prior knowledge based, i.e., it is a deductive inference system non-adaptive to input data rather than an inductive system capable of learning from data (refer to Section II-B), SIAM™ requires no reference data sample to run. It also requires no human assistance to define system-free parameters based on heuristics, like its alternative statistical approaches (see Table I). Thus, the SIAM™ cost in manpower is equal to zero (refer to this section above). Its costs in terms of computer power are almost negligible (refer to this section above).

VIII. New Inter-discplinary Research and Market Opportunities

Based on the exiting literature in combination with experimental evidence collected in this work, the operational, automatic, near real-time, multi-sensor, multi-resolution SIAM™ appears eligible for opening up new inter-disciplinary research and market opportunities, such as those listed below, in compliance with the visionary goal of the GEOSS initiative and the QA4EO guidelines [3] (refer to Section I).

1) Improve the OQIs of existing commercial RS-IUS software products such as those listed in Table I, including state-of-the-art two-stage non-iterative GEOBIA and three-stage iterative GEOOIA commercial software products (refer to Section II-F). For example, SIAM™ would be eligible for use as pre-attentive vision first stage in operational multi-sensor, multi-resolution, MS RS-IUSs provided with a feedback mechanism to enhance the RS image pre-processing phase (e.g., through stratified TOC [10]), refer to the existing literature [5]–[17], to the ATCOR-2/3/4 data workflow [83]–[86] shown in Fig. 1 and to Section II-G.
2) Automatic transformation of sub-symbolic, raster EO data into symbolic, vector, geospatial information made available in a GIS-ready format. In other words, SIAM™ provides seamless integration of RS imagery with GI-Science [25], [141] (geomatics engineering [69]), refer to Section II-F.
3) Integration of Internet-based satellite mapping-on-demand with web-based GIS and virtual earth geo-browsers such as the hugely popular Google Earth, NASA's World Wind, and Microsoft Virtual Earth, see Fig. 16.

4) Development of semantic querying systems of large-scale multi-source RS image databases where SIAM™ can be exploited as an automatic source of reference classification maps. This would represent a dramatic improvement over non-semantic query modes currently available in image database retrieval systems based on text-driven query strategies and query by either an image, object or multi-object example.
5) Development of so-called fourth generation FIEOSs ([137]) where SIAM™ can be mounted on board. The same consideration holds for ground receiving stations which could be provided with an operational automatic "intelligent" data processing feedback system.
6) Dissemination of advanced EO expertise, science, and technology in developing countries and emerging countries. Automatic EO image understanding technologies are "democratic" in nature, i.e., eligible for use by all. In other words, EO researchers and institutions should perceive SIAM™ as a novel technical opportunity to pursue ethical issues.

IX. Conclusion

This original work presents a novel (to the best of these authors' knowledge, the first) probability sampling protocol for thematic and spatial quality assessments of thematic maps generated from spaceborne/airborne VHR images. The proposed protocol delivers as output an original set of mutually uncorrelated TQIs and SQIs featuring:

- Statistical consistency (validity), i.e., sample estimates are provided with the necessary probability foundation to permit generalization from the sample data subset to the whole target population being sampled [55], [60].
- Statistical significance, i.e., TQIs and SQIs are provided with a degree of uncertainty in measurement in compliance with the principles of statistics together with the QA4EO international guidelines [3].

Independent of TQIs and SQIs, an original CVPSI is estimated as a fuzzy degree of match between a reference and a test semantic vocabulary, which may not coincide.

The proposed protocol is validated in the quality assessment of preliminary classification maps automatically generated from VHR optical images by the operational, near real-time SIAM™ software product [5]–[17]. Provided by DigitalGlobe for testing purposes, the VHR image set consists of two WV-2 images of the city area of Brazilia (Brazil), acquired in the wet (T1) and dry (T2) seasons of year 2010, and one QB-2 image of Brazilia, acquired at time $T1 + 45$ days.

Although underestimated in RS common practice, radiometric calibration data pre-processing is considered: 1) critical to RS data QA and, therefore, data usability, in compliance with the QA4EO guidelines [3], and 2) a necessary, although not sufficient, condition to automate a satellite-based information processing system, like SIAM™ [7]. In this paper, the three VHR test images are radiometrically calibrated into TOARF values. In addition, to reduce the inherent data spread due to varying acquisition conditions (e.g., sensor viewing position, Sun position, atmospheric conditions, etc.) over the same surface type, the two "slave," off-nadir, 2-m resolution WV-2 images in TOARF values are radiometrically registered to a pair of "master," nadir-view, 30-m resolution Landsat-7 ETM+ images radiometrically calibrated into TOARF values.

The main experimental conclusion of this work is that the proposed protocol is tested successfully in the accuracy validation of the Q-SIAM™ multi-granularity maps automatically generated from multi-sensor multi-temporal VHR images. In these experiments, collected TQIs and SQIs are statistically valid, statistically significant, and consistent across different thematic maps; they comply with theoretical considerations and agree with visual (qualitative) evidence, collected CVPSI values and (quantitative) QIs of operativeness (OQIs) claimed for SIAM™ by the existing literature [5]–[17]. Estimated SQIs are found to be biased due to a summation of effects. First, an eight-adjacency neighborhood phenomenon causes oversegmentation quality, see (28), to be overestimated while undersegmentation quality, see (29), and edge position quality, see (30) and (31), are underestimated. Second, an inadequacy to cope with a test and a reference semantic vocabulary when they do not coincide causes all aforementioned SQIs to be underestimated.

As a subsidiary conclusion, the statistically consistent and statistically significant accuracy validation of the Q-SIAM™ maps accomplished in this work, together with OQIs claimed for SIAM™ by existing literature [5]–[17], makes the operational (fast, accurate, automatic, robust, scalable) SIAM™ preliminary classification software product eligible for opening up new inter-disciplinary research and market opportunities in accordance with the visionary goal of the GEOSS initiative and the QA4EO guidelines [3].

Appendix

The probability $p(T)$ of selecting a target element T out of N units, where N is the population size, in a sequence of n independent yes/no experiments, each of which yields success with probability $p(1/N)$, is equal to 1 minus the probability of selecting all the remaining non-target (NT) elements, $p(NT)$, across the n independent yes/no experiments, i.e.,

$$p(T) = 1 - p(NT) \tag{33}$$

where

$$p(NT) = p(NT_1, \ldots, NT_n) = \prod_i p(NT_i) = (N-1)^n / N^n. \tag{34}$$

Thus

$$p(T) = 1 - (N-1)^n / N^n \tag{35}$$

such that $p(T) \to 1$ if $n \to \infty$, i.e., the inclusion probability of the target element T tends to 1 as the number of experiments goes to infinity.

According to the binomial expansion:

$$(a-b)^n = \sum_{k=0}^{n} \frac{n!}{(n-k)!k!} a^{n-k} b^k (-1)^k. \tag{36}$$

Thus, if $a = N =$ size of the population to be sampled, $b = 1 =$ number of target elements in the population to be sampled, $n =$ number of experiments where one single sample is selected, then the inclusion probability of that target element in n experiments is:

$$\begin{aligned} p(T) &= 1 - \frac{(N-1)^n}{N^n} = \frac{N^n - (N-1)^n}{N^n} \\ &= \frac{N^n - \sum_{k=0}^{n} \frac{n!}{(n-k)!k!} N^{n-k}(-1)^k}{N^n} \\ &= \frac{N^n - N^n + nN^{n-1} - \frac{n(n-1)}{2!}N^{n-2} - \ldots - (-1)^n}{N^n} \end{aligned} \tag{37}$$

hence, if the population size $N \to \infty$, then (37) becomes

$$\begin{aligned} p(T) &= \frac{nN^{n-1} - \frac{n(n-1)}{2!}N^{n-2} + \ldots + (-1)^n}{N^n} \\ &\approx \frac{nN^{n-1}}{N^n} = \frac{n}{N}. \end{aligned}$$

Thus, (11) is proved.

ACKNOWLEDGMENT

These authors wish to thank I. Gilbert of DigitalGlobe who provided the WV-2 and QB-2 images employed in this research. A. Baraldi thanks R. Capurro for his hospitality, patience, politeness, and open-mindedness. The authors acknowledge intellectual debt to S. V. Stehman, College of Environmental Science and Forestry, State University of New York, whose published papers and helpful comments inspired our work. The authors also wish to thank the Editor-in-Chief, Associate Editor, and reviewers for their competence, patience, and willingness to help.

REFERENCES

[1] O. Sjahputera, C. H. Davis, B. Claywell, N. J. Hudson, J. M. Keller, M. G. Vincent, Y. Li, M. Klaric, and C. R. Shyu, "GeoCDX: An automated change detection and exploitation system for high resolution satellite imagery," in *Proc. IEEE IGARSS*, Boston, MA, USA, Jul. 6–11, 2008, pp. 467–470.

[2] G. Gutman, A. C. Janetos, C. O. Justice, E. F. Moran, J. F. Mustard, R. R. Rindfuss, D. Skole, B. L. Turner, and M. A. Cochrane, Eds., *Land Change Science*. Dordrecht, The Netherlands: Kluwer, 2004.

[3] *A Quality Assurance Framework for Earth Observation*, ver. 4.0, GEO/CEOSS, Geneva, Switzerland, Jan. 2010. [Online]. Available: http://qa4eo.org/docs/QA4EO_Principles_v4.0.pdf

[4] GEO. (2005, Feb. 16). The Global Earth Observation System of Systems (GEOSS) 10-Year Implementation Plan, Geneva, Switzerland, (accessed on 10 January 2012). [Online]. Available: http://www.earthobservations.org/docs/10-Year%20Implementation%20Plan.pdf

[5] A. Baraldi, "Impact of radiometric calibration and specifications of spaceborne optical imaging sensors on the development of operational automatic remote sensing image understanding systems," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 2, no. 2, pp. 104–134, Jun. 2009.

[6] A. Baraldi, V. Puzzolo, P. Blonda, L. Bruzzone, and C. Tarantino, "Automatic spectral rule-based preliminary mapping of calibrated Landsat TM and ETM+ images," *IEEE Trans. Geosci. Remote Sens.*, vol. 44, no. 9, pp. 2563–2586, Sep. 2006.

[7] A. Baraldi, L. Durieux, D. Simonetti, G. Conchedda, F. Holecz, and P. Blonda, "Automatic spectral rule-based preliminary classification of radiometrically calibrated SPOT-4/-5/IRS, AVHRR/MSG, AATSR, IKONOS/QuickBird/OrbView/GeoEye and DMC/SPOT-1/-2 imagery—Part I: System design and implementation," *IEEE Trans. Geosci. Remote Sens.*, vol. 48, no. 3, pp. 1299–1325, Mar. 2010.

[8] A. Baraldi, L. Durieux, D. Simonetti, G. Conchedda, F. Holecz, and P. Blonda, "Automatic spectral rule-based preliminary classification of radiometrically calibrated SPOT-4/-5/IRS, AVHRR/MSG, AATSR, IKONOS/QuickBird/OrbView/GeoEye and DMC/SPOT-1/-2 imagery—Part II: Classification accuracy assessment," *IEEE Trans. Geosci. Remote Sens.*, vol. 48, no. 3, pp. 1326–1354, Mar. 2010.

[9] A. Baraldi, L. Durieux, D. Simonetti, G. Conchedda, F. Holecz, and P. Blonda, "Corrections to Automatic spectral rule-based preliminary classification of radiometrically calibrated SPOT-4/-5/IRS, AVHRR/MSG, AATSR, IKONOS/QuickBird/OrbView/GeoEye, and DMC/SPOT-1/-2 imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 48, no. 3, p. 1635, Mar. 2010.

[10] A. Baraldi, M. Gironda, and D. Simonetti, "Operational two-stage stratified topographic correction of spaceborne multi-spectral imagery employing an automatic spectral rule-based decision-tree preliminary classifier," *IEEE Trans. Geosci. Remote Sens.*, vol. 48, no. 1, pp. 112–146, Jan. 2010.

[11] A. Baraldi, T. Wassenaar, and S. Kay, "Operational performance of an automatic preliminary spectral rule-based decision-tree classifier of spaceborne very high resolution optical images," *IEEE Trans. Geosci. Remote Sens.*, vol. 48, no. 9, pp. 3482–3502, Sep. 2010.

[12] A. Baraldi, "Fuzzification of a crisp near-real-time operational automatic spectral-rule-based decision-tree preliminary classifier of multisource multispectral remotely sensed images," *IEEE Trans. Geosci. Remote Sens.*, vol. 49, no. 6, pp. 2113–2134, Jun. 2011.

[13] A. Baraldi, "Vision goes symbolic without loss of information within the preattentive vision phase: The need to shift the learning paradigm from Machine-Learning (from examples) to Machine-Teaching (by rules) at the first stage of a two-stage hybrid remote sensing image understanding system, Part I: Introduction," in *Earth Observation*. Rijeka, Croatia: InTech, 2012, pp. 63–98.

[14] A. Baraldi, "Vision goes symbolic without loss of information within the preattentive vision phase: The need to shift the learning paradigm from Machine-Learning (from examples) to Machine-Teaching (by rules) at the first stage of a two-stage hybrid remote sensing image understanding system, Part II: Introduction," in *Earth Observation*. Rijeka, Croatia: InTech, 2012, pp. 99–136.

[15] A. Baraldi, "Satellite image automatic mapper (SIAM™)—A turnkey software executable for automatic near real-time multi-sensor multi-resolution spectral rule-based preliminary classification of spaceborne multi-spectral images," *Recent Patents Space Technol.*, vol. 1, no. 2, pp. 81–106, Dec. 2011.

[16] A. Baraldi and L. Boschetti, "Operational automatic remote sensing image understanding systems: Beyond Geographic Object-Based and Object-Oriented Image Analysis (GEOBIA/GEOOIA)—Part 1: Introduction," *Remote Sens.*, vol. 4, no. 9, pp. 2694–2735, Sep. 2012.

[17] A. Baraldi and L. Boschetti, "Operational automatic remote sensing image understanding systems: Beyond Geographic Object-Based and Object-Oriented Image Analysis (GEOBIA/GEOOIA)—Part 2: Novel system architecture, information/knowledge representation, algorithm design and implementation," *Remote Sens.*, vol. 4, no. 9, pp. 2768–2817, Sep. 2012.

[18] P. Zamperoni, "Plus ça va, moins ça va," *Pattern Recognit. Lett.*, vol. 17, no. 7, pp. 671–677, Jun. 1996.

[19] E. Diamant, "Machine Learning: When and Where the Horses Went Astray?" in *Machine Learning*, Y. Zhang, Ed. Rijeka, Croatia: InTech, 2010, pp. 1–18.

[20] M. Sonka, V. Hlavac, and R. Boyle, *Image Processing, Analysis and Machine Vision*. London, U.K.: Chapman & Hall, 1994.

[21] D. Marr, *Vision*. New York, NY, USA: Freeman, 1982.

[22] F. Wang, "Fuzzy supervised classification of remote sensing images," *IEEE Trans. Geosci Remote Sens.*, vol. 28, no. 2, pp. 194–201, Mar. 1990.

[23] R. Capurro and B. Hjørland, *The concept of information, Annual Review of Information Science and Technology*, B. Cronin, Ed., Medford, NJ, USA: Information Today Inc., 2003, vol. 37, ch. 8, pp. 343–411. [Online]. Available: http://www.capurro.de/infoconcept.html

[24] R. Capurro, *Hermeneutics and the Phenomenon of Information., in Metaphysics, Epistemology, and Technology. Research in Philosophy and Technology*, vol. 19. Amsterdam, The Netherlands: Elsevier, 2000, pp. 79–85.

[25] M. F. Goodchild, M. Yuan, and T. J. Cova, "Towards a general theory of geographic representation in GIS," *Int. J. Geogr. Inf. Sci.*, vol. 21, no. 3, pp. 239–260, Jan. 2007.

[26] A. D. Tonchev and C. D. Tonchev, *Method for Measuring the Overall Operational Performance of Hydrocarbon Facilities*, Patent 20 080 262 898, Oct. 23, 2008. [Online]. Available: http://www.faqs.org/patents/app/20080262898

[27] C. Mason and E. R. Kandel, "Central Visual Pathways," in *Principles of Neural Science*, E. Kandel and J. Schwartz, Eds. Norwalk, CT, USA: Appleton and Lange, 1991, pp. 420–439.

[28] P. Gouras, "Color Vision," in *Principles of Neural Science*, E. Kandel and J. Schwartz, Eds. Norwalk, CT, USA: Appleton and Lange, 1991, pp. 467–479.

[29] E. R. Kandel, "Perception of Motion, Depth and Form," in *Principles of Neural Science*, E. Kandel and J. Schwartz, Eds. Norwalk, CT, USA: Appleton and Lange, 1991, pp. 441–466.

[30] H. R. Wilson and J. R. Bergen, "A four mechanism model for threshold spatial vision," *Vis. Res.*, vol. 19, no. 1, pp. 19–32, 1979.

[31] S. P. Vecera and M. J. Farah, "Is visual image segmentation a bottom-up or an interactive process?" *Percept. Psychophys.*, vol. 59, no. 8, pp. 1280–1296, Nov. 1997.

[32] V. Cherkassky and F. Mulier, *Learning from Data: Concepts, Theory, and Methods*. New York, NY, USA: Wiley, 1998.

[33] P. Mather, *Computer Processing of Remotely-Sensed Images—An Introduction*. Chichester, U.K.: Wiley, 1994.

[34] C. M. Bishop, *Neural Networks for Pattern Recognition*. Oxford, U.K.: Clarendon, 1995.

[35] T. Matsuyama and V. Shang-Shouq Hwang, *SIGMA—A Knowledge-Based Aerial Image Understanding System*. New York, NY, USA: Plenum, 1990.

[36] K. Pakzad, J. Bäckner, and S. Growe, Knowledge based moorland interpretation using a hybrid system for image analysis," in *Proc. ISPRS Conf.*, Munich, Germany, Sep. 8–10, 1999, pp. 1103–1110. [Online]. Available: http://www.tnt.uni-hannover.de/papers/view.php?ind=1999&ord=Authors&mod=ASC

[37] S. Growe, "Knowledge based interpretation of multisensor and multi-temporal remote sensing images," *Int. Arch. Photogramm. Remote Sens.*, vol. 32, pt. 7-4-3 W6, pp. 130–138, Jun. 1999.

[38] M. Baatz and A. Schäpe, "Multi-resolution Segmentation: An optimization approach for high quality multi-scale image segmentation," *J. Photogramm. Remote Sens.*, vol. 58, no. 3, pp. 12–23, 2000.

[39] G. J. Hay and G. Castilla, "Object-based image analysis: Strengths, Weaknesses, Opportunities and Threats (SWOT)," in *Proc. 1st Int. Conf. OBIA*, Salzburg, Austria, Jul. 4–5, 2006, pp. 1–3.

[40] G. J. Hay and G. Castilla, "Geographic object-based image analysis (GEOBIA): A new name for a new discipline," in *Object-Based Image Analysis: Spatial Concepts for Knowledge-driven Remote Sensing Applications*, T. Blaschke, S. Lang, and G. J. Hay, Eds. New York, NY, USA: Springer-Verlag, 2008, ch. 1.4, pp. 81–92.

[41] J. Marroquin, S. Mitter, and T. Poggio, "Probabilistic solution of ill-posed problems in computational vision," *J. Amer. Stat. Assoc.*, vol. 82, no. 397, pp. 76–89, Mar. 1987.

[42] M. Bertero, T. Poggio, and V. Torre, "Ill-posed problems in early vision," *Proc. IEEE*, vol. 76, no. 8, pp. 869–889, Aug. 1988.

[43] P. Corcoran and A. Winstanley, "Using Texture to Tackle the Problem of Scale in Landcover Classification," in *Object-Based Image Analysis: Spatial Concepts for Knowledge-Driven Remote Sensing Applications*, T. Blaschke, S. Lang, and G. J. Hay, Eds. New York, NY, USA: Springer-Verlag, 2008, pp. 113–132.

[44] M. Petrou and P. Sevilla, *Image Processing: Dealing with Texture*. Chichester, U.K.: Wiley, 2006.

[45] D. C. Burr and M. C. Morrone, "A Nonlinear Model of Feature Detection," in *Nonlinear Vision: Determination of Neural Receptive Fields, Functions, and Networks*, R. B. Pinter and N. Bahram, Eds. Boca Raton, FL, USA: CRC Press, 1992, pp. 309–327.

[46] J. Hadamard, "Sur les problemes aux derivees partielles et leur signification physique," *Princeton Univ. Bull.*, vol. 13, pp. 49–52, 1902.

[47] Q. Iqbal and J. K. Aggarwal, "Image retrieval via isotropic and anisotropic mappings," in *Proc. IAPR Workshop Pattern Recognit. Inf. Syst.*, Setubal, Portugal, Jul. 6–8, 2001, pp. 34–49, 2001.

[48] CEOS Working Group on Calibration and Validation , Land Product Validation Subgroup, accessed on 10 January 2012. [Online]. Available: http://lpvs.gsfc.nasa.gov/

[49] S. V. Stehman, "Selecting and interpreting measures of thematic classification accuracy," *Remote Sens. Environ.*, vol. 62, no. 1, pp. 77–89, Oct. 1997.

[50] R. S. Lunetta and C. D. Elvidge, *Remote sensing and Change Detection: Environmental Monitoring Methods and Applications*. Chelsea, MI, USA: Ann Arbor Press, 1998.

[51] R. G. Congalton and K. Green, *Assessing the Accuracy of Remotely Sensed Data*. Boca Raton, FL, USA: Lewis Publishers, 1999.

[52] A. Baraldi, L. Bruzzone, and P. Blonda, "Quality assessment of classification and cluster maps without ground truth knowledge," *IEEE Trans. Geosci. Remote Sens.*, vol. 43, no. 4, pp. 857–873, Apr. 2005.

[53] G. M. Foody, "Status of land cover classification accuracy assessment," *Remote Sens. Environ.*, vol. 80, no. 1, pp. 185–201, Apr. 2002.

[54] C. Persello and L. Bruzzone, "A novel protocol for accuracy assessment in classification of very high resolution images," *IEEE Trans. Geosci. Remote Sens.*, vol. 48, no. 3, pp. 1232–1244, Mar. 2010.

[55] S. V. Stehman and R. L. Czaplewski, "Design and analysis for thematic map accuracy assessment: Fundamental principles," *Remote Sens. Environ.*, vol. 64, no. 3, pp. 331–344, Jun. 1998.

[56] R. G. Pontius, "Quantification error versus location error in comparison of categorical maps," *Photogramm. Eng. Remote Sens.*, vol. 66, no. 8, pp. 1011–1016, Aug. 2000.

[57] L. Wald, T. Ranchin, and M. Mangolini, "Fusion of satellite images of different spatial resolutions: Assessing the quality of resulting images," *Photogramm. Eng. Remote Sens.*, vol. 63, no. 6, pp. 691–699, Jun. 1997.

[58] Z. Wang and A. C. Bovik, "A universal image quality index," *IEEE Signal Process. Lett.*, vol. 9, no. 3, pp. 81–84, Mar. 2002.

[59] L. Alparone, S. Baronti, A. Garzelli, and F. Nencini, "A global quality measurement of pan-sharpened multispectral imagery," *IEEE Geosci. Remote Sens. Lett.*, vol. 1, no. 4, pp. 313–317, Oct. 2004.

[60] W. S. Overton and S. V. Stehman, "The Horvitz–Thompson theorem as a unifying perspective for probability sampling: With examples from natural resource sampling," *Amer. Stat.*, vol. 49, no. 3, pp. 261–268, Aug. 1995.

[61] F. Van Coillie, R. Pires, N. Van Camp, and S. Gautama, "Quantitative segmentation evaluation for large scale mapping purposes," *Int. Arch. Phtotogramm., Remote Sens. Spatial Inf. Sci.*, vol. 38, no. 4/C1, p. 6, 2008 (accessed on 10 Jan. 2012). [Online]. Available: https://biblio.ugent.be/publication/854114

[62] F. Van Coillie, N. Van Camp, R. De Wulf, L. Bral, and S. Gautama, "Segmentation quality evaluation for large scale mapping purposes in Flanders, Belgium," in *Proc. Int. Arch. Phtotogramm., Remote Sens. Spatial Inf. Sci.*, 2010, pp. 1–4. [Online]. Available: https://biblio.ugent.be/publication/1220278

[63] S. V. Stehman, "Comparing thematic maps based on map value," *Int. J. Remote Sens.*, vol. 20, no. 12, pp. 2347–2366, Jan. 1999.

[64] L. A. Zadeh, "Fuzzy sets," *Inf. Control*, vol. 8, no. 3, pp. 338–353, Jun. 1965.

[65] B. Kosko, *Fuzzy Thinking*. London, U.K.: Flamingo, 1994.

[66] K. Kuzera and R. G. Pontius, "Importance of matrix construction for multiple-resolution categorical map comparison," *GISci. Remote Sens.*, vol. 45, no. 3, pp. 249–274, 2008.

[67] M. Beauchemin and K. Thomson, "The evaluation of segmentation results and the overlapping area matrix," *Int. J. Remote Sens.*, vol. 18, no. 18, pp. 3895–3899, Dec. 1997.

[68] O. Ahlqvist, "Extending post-classification change detection using semantic similarity metrics to overcome class heterogeneity: A study of 1992 and 2001 U.S. National Land Cover Databases changes," *Remote Sens. Environ.*, vol. 112, no. 3, pp. 1226–1241, Mar. 2008.

[69] R. Laurini and D. Thompson, *Fundamentals of Spatial Information Systems*. San Diego, CA, USA: Academic, 1992.

[70] H. Mizen, C. Dolbear, and G. Hart, "Ontology ontogeny: Understanding how an ontology is created and developed," in *Proc. 1st Int. Conf. GeoS*, Mexico City, Mexico, Nov. 29/30, 2005, pp. 15–29.

[71] F. T. Fonseca, M. J. Egenhofer, P. Agouris, and G. Camara, "Using ontologies for integrated geographic information systems," *Trans. GIS*, vol. 6, no. 3, pp. 231–257, Jun. 2002.

[72] N. Guarino, "Formal ontology, conceptual analysis and knowledge representation," *Int. J. Hum. Comput. Studies*, vol. 43, no. 5/6, pp. 625–640, Nov./Dec. 1995.

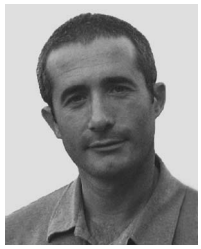
[73] J. F. Sowa, *Knowledge representation: Logical, philosophical, and computational foundations*. Cambridge, MA, USA: MIT Press, 2000.

[74] O. Ahlqvist, "Using uncertain conceptual spaces to translate between land cover categories," *Int. J. Geogr. Inf. Sci.*, vol. 19, no. 7, pp. 831–857, Aug. 2005.

[75] C. C. Feng and D. M. Flewelling, "Assessment of semantic similarity between land use/land cover classification systems," *Comput., Environ. Urban Syst.*, vol. 28, no. 3, pp. 229–246, May 2004.

[76] M. Kavouras and M. Kokla, "A method for the formalization and integration of geographical categorizations," *Int. J. Geogr. Inf. Sci.*, vol. 16, no. 5, pp. 439–453, Jul. 2002.
[77] A. K. Shackelford and C. H. Davis, "A hierarchical fuzzy classification approach for high-resolution multispectral data over urban areas," *IEEE Trans. Geosci. Remote Sens.*, vol. 41, no. 9, pp. 1920–1932, Sep. 2003.
[78] C. Shannon, "A mathematical theory of communication," *Bell Syst. Tech. J.*, vol. 27, pp. 379–423, Jul.–Oct. 1948.
[79] S. Lang, "Object-based image analysis for remote sesning applications: Modeling reality—Dealing with complexity," in *Object-Based Image Analysis-Spatial Concepts for Knowledge-driven Remote Sensing Applications*, T. Blaschke, S. Lang, and G. J. Hay, Eds. New York, NY, USA: Springer-Verlag, 2008, ch. 1.1, pp. 3–27.
[80] P. Läscher, D. Burghardt, and R. Weibel, "Ontology-driven enrichment of spatial databases," in *Proc. 10th ICA Workshop Gen. Multiple Represent.*, Moscow, Russia, Aug. 2/3, 2007, pp. 1–13. [Online]. Available: http://www.geo.uzh.ch/~luescher/publications/luescher_genws2007.pdf
[81] G. Schaepman-Strub, M. E. Schaepman, T. H. Painter, S. Dangel, and J. V. Martonchik, "Reflectance quantities in optical remote sensing-definitions and case studies," *Remote Sens. Environ.*, vol. 103, no. 1, pp. 27–42, Jul. 2006.
[82] M. Herold, C. Woodcock, A. Di Gregorio, P. Mayaux, A. S. Belward, J. Latham, and C. Schmullius, "A joint initiative for harmonization and validation of land cover datasets," *IEEE Trans. Geosci. Remote Sens.*, vol. 44, no. 7, pp. 1719–1727, Jul. 2006.
[83] R. Richter and D. Schläpfer, *Atmospheric/Topographic Correction for Satellite Imagery—ATCOR-2/3 User Guide*, ReSe Appl. Schläpfer, Wil, Switzerland, Version 8.0.2. [Online]. Available: http://www.rese.ch/pdf/atcor3_manual.pdf
[84] R. Richter and D. Schläpfer, *Atmospheric/Topographic Correction for Airborne Imagery—ATCOR-4 User Guide*, ReSe Appl. Schläpfer, Wil, Switzerland, Version 6.2 BETA. [Online]. Available: http://www.dlr.de/eoc/Portaldata/60/Resources/dokumente/5_tech_mod/atcor4_manual_2012.pdf
[85] W. Dorigo, R. Richter, R. F. Baret, R. Bamler, and W. Wagner, "Enhanced automated canopy characterization from hyperspectral data by a novel two step radiative transfer model inversion approach," *Remote Sens.*, vol. 1, no. 4, pp. 1139–1170, Nov. 2009.
[86] D. Schläpfer, R. Richter, and A. Hueni, "Recent developments in operational atmospheric and radiometric correction of hyperspectral imagery," in *Proc. 6th EARSeL SIG IS Workshop*, Tel-Aviv, Israel, Mar. 16–19, 2009, pp. 1–7.
[87] S. Liang, *Quantitative Remote Sensing of Land Surfaces*. Hoboken, NJ, USA: Wiley, 2004.
[88] D. Maniates and D. Mollicone, "Options for sampling and stratification for national forest inventories to implement REDD+ under the UNFCCC," *Carbon Bal. Manage.*, vol. 5, no. 9, pp. 1–14, Dec. 2010.
[89] *eCognition Elements User Guide 4*, Definiens Imaging GmbH, Munich, Germany, 2004.
[90] *Developer 8 Reference Book*, Definiens AG, Munich, Germany, 2011.
[91] T. Esch, M. Thiel, M. Bock, A. Roth, and S. Dech, "Improvement of image segmentation accuracy based on multiscale optimization procedure," *IEEE Geosci. Remote Sens. Lett.*, vol. 5, no. 3, pp. 463–467, Jul. 2008.
[92] M. Nuebert, H. Herold, and G. Meinel, "Assessing image segmentation quality—Concepts, methods and application," in *Object-Based Image Analysis-Spatial Concepts for Knowledge-driven Remote Sensing Applications*, T. Blaschke, S. Lang, and G. J. Hay, Eds. New York, NY, USA: Springer-Verlag, 2008, ch. 8.3, pp. 769–784.
[93] P. S. Chavez, "An improved dark-object subtraction technique for atmospheric scattering correction of multispectral data," *Remote Sens. Environ.*, vol. 24, no. 3, pp. 459–479, Apr. 1988.
[94] P. H. Swain and S. M. Davis, *Remote Sensing: The Quantitative Approach*. New York, NY, USA: McGraw Hill, 1978.
[95] Group on Earth Observations. (2008, Nov.). GEO Announces Free and Unrestricted Access to Full Landsat Archive: Universal Availability of Cost-Free Satellite Data and Images will Revolutionize The Use of Earth Observations for Decision-Making, Geneva, Switzerland. [Online]. Available: www.fabricadebani.ro/userfiles/GEO_press_release.doc
[96] ESA, (accessed 9 Sep. 2012) About GMES—Overview. [Online]. Available: http://www.esa.int/esaLP/SEMRRI0DU8E_LPgmes_0.html
[97] GMES, (accessed 10 Jan. 2012) GMES Info. [Online]. Available: http://www.gmes.info
[98] USGS, (accessed 9 Sep. 2012) Web-Enabled Landsat Data (WELD) Project. [Online]. Available: http://landsat.usgs.gov/WELD.php
[99] A. K. Shackelford and C. H. Davis, "A combined fuzzy pixel-based and object-based approach for classification of high-resolution multispectral data over urban areas," *IEEE Trans. Geosci. Remote Sens.*, vol. 41, no. 10, pp. 2354–2363, Oct. 2003.
[100] A. Baraldi and F. Parmiggiani, "Urban area classification by multispectral SPOT images," *IEEE Trans. Geosci. Remote Sens.*, vol. 28, no. 4, pp. 674–680, Jul. 1990.
[101] Y. I. Zhang, "A survey on evaluation methods for image segmentation," *Pattern Recognit.*, vol. 29, no. 8, pp. 1335–1346, Aug. 1996.
[102] H. Zhang, J. Fritts, and S. Goldman, "Image segmentation evaluation: A survey of unsupervised methods," *Comput. Vis. Image Understand.*, vol. 110, no. 2, pp. 260–280, May 2008.
[103] G. M. Foody, "Thematic map comparison: Evaluating the statistical significance of differences in classification accuracy," *Photogramm. Eng. Remote Sens.*, vol. 70, no. 5, pp. 627–633, May 2004.
[104] A. Wu, Z. Li, and J. Cihlar, "Effects of land cover type and greenness on advanced very high resolution radiometer bidirectional reflectances: Analysis and removal," *J. Geophys. Res.*, vol. 100, no. D5, pp. 9179–9192, Jan 1995.
[105] J. D. Shepherd and J. R. Dymond, "BRDF correction of vegetation in AVHRR imagery," *Remote Sens. Environ.*, vol. 74, no. 3, pp. 397–408, Dec. 2000.
[106] T. Danaher, "An empirical BRDF correction for landsat TM and ETM+ imagery," in *Proc. 11th Aust. Remote Sens. Photogramm. Conf.*, Brisbane, Australia, 2002, pp. 2654–2657.
[107] E. Vermote, D. Tanré, J. L. Deuzé, M. Herman, and J. J. Morcrette, Second Simulation of the Satellite Signal in the Solar Spectrum (6S), 6S User Guide Version 2, Lab. Opt. Atmos., Villeneuve d'Ascq, France, (accessed 10 Jan. 2012). [Online]. Available: http://www.rsgis.ait.ac.th/~honda/textbooks/advrs/6smanv2.0_P1.pdf
[108] M. P. Bishop and J. D. Colby, "Anisotropic reflectance correction of SPOT-3 HRV imagery," *Int. J. Remote Sens.*, vol. 23, no. 10, pp. 2125–2131, May 2002.
[109] M. P. Bishop, J. F. Shroder, and J. D. Colby, "Remote sensing and geomorphometry for studying relief production in high mountains," *Geomorphology*, vol. 55, no. 1–4, pp. 345–361, Sep. 2003.
[110] DigitalGlobe. Radiometric Use of QuickBird Imagery, Longmont, CO, USA, (accessed 10 Jan. 2012). [Online]. Available: http://www.digitalglobe.com/downloads/QuickBird_technote_raduse_v1.pdf
[111] T. Lillesand and R. Kiefer, *Remote Sensing and Image Interpretation*. New York, NY, USA: Wiley, 1979.
[112] L. Leigh, D. Helder, I. Behnert, A. Deadman, N. Fox, U. M. Leloglu, H. Ozen, and D. Griffith, "Tuz Gölä site characteristics," in *Proc. IEEE IGARSS*, Jul. 24–29, 2011, pp. 3871–3874.
[113] F. Li, D. L. B. Jupp, and M. Thankappana, "Using high resolution DSM data to correct the terrain illumination effect in Landsat data," in *Proc. 19th Int. Congr. Modelling Simul.*, Perth, Australia, Dec. 12–16, 2011, pp. 2402–2408.
[114] X. Wu, S. Collings, and P. Caccetta, "BRDF and illumination calibration for very high resolution imaging sensors," in *Proc. IEEE IGARSS*, Jul. 25–30, 2010, pp. 3162–3165.
[115] M. Humber, A. Baraldi, and L. Boschetti, "Automatic near real-time preliminary classification of spaceborne/airborne multi-spectral images: Accuracy validation of the Satellite Image Automatic Mapper (SIAM™) and Atmospheric/Topographic Correction (ATCOR)-Spectral Classification (SPECL) software products in operating mode," *Remote Sensing*, 2013.
[116] (accessed on 1 Nov. 2010). [Online]. Available: http://www.digitalglobe.com/index.php/70/Product+Samples
[117] T. Novack, T. Esch, H. Kux, and U. Stilla, "Machine learning comparison between WorldView-2 and QuickBird-2 simulated imagery regarding object-based urban land cover classification," *Remote Sens.*, vol. 3, no. 10, pp. 2263–2282, Oct. 2011.
[118] C. Tarantino, M. Adamo, G. Pasquariello, F. Lovergine, P. Blonda, and V. Tomaselli, "8-band image data processing of the WorldView-2 satellite in a wide area of applications," in *Earth Observation*. Rijeka, Croatia: InTech, 2012, pp. 137–152.
[119] H. Z. Mohd Shafri, M. A. Mohd Salleh, and A. Ghiyamat, "Hyperspectral remote sensing of vegetation using red edge position techniques," *Amer. J. Appl. Sci.*, vol. 3, no. 6, pp. 1864–1871, Jun. 2006.
[120] M. Boschetti, L. Boschetti, S. Oliveri, L. Casati, and I. Canova, "Tree species mapping with Airborne hyper-spectral MIVIS data: The Ticino Park study case," *Int. J. Remote Sens.*, vol. 28, no. 6, pp. 1251–1261, Mar. 2007.
[121] K. Richter, T. B. Hank, F. Vuolo, W. Mauser, and G. D'Urso, "Optimal exploitation of the Sentinel-2 spectral capabilities for crop leaf area index mapping," *Remote Sens.*, vol. 4, no. 3, pp. 561–582, Feb. 2012.
[122] G. Chander, B. L. Markham, and D. L. Helder, "Summary of current radiometric calibration coefficients for Landsat MSS, TM, ETM+, and

EO-1 ALI sensors," *Remote Sens. Environ.*, vol. 113, no. 5, pp. 893–903, May 2009.
[123] S. V. Stehman and J. D. Wickham, "Pixels, blocks of pixels, and polygons: Choosing a spatial unit for thematic accuracy assessment," *Remote Sens. Environ.*, vol. 115, no. 12, pp. 3044–3055, Dec. 2011.
[124] J. Brinkworth, *Software Quality Management*. New York, NY, USA: Prentice-Hall, 1992.
[125] F. Salge, "Semantic accuracy," in *Elements of Spatial Data Quality*, S. C. Guptill and J. L. Morisson, Eds. Oxford, U.K.: Elsevier, 1995, pp. 139–151.
[126] Y. Bishr and Y. , "Overcoming the semantic and other barriers to GIS interoperability," *Int. J. Geogr. Inf. Sci.*, vol. 12, no. 4, pp. 299–314, Jun. 1998.
[127] R. G. Pontius, Jr. and M. Millones, "Death to Kappa: Birth of quantity disagreement and allocation disagreement for accuracy assessment," *Int. J. Remote Sens.*, vol. 32, no. 15, pp. 4407–4429, Aug. 2011.
[128] R. G. Pontius and J. Connors, "Expanding the conceptual, mathematical and practical methods for map comparison," in *Proc. Conf. Spatial Accuracy*, Lisbon, Portugal, 2006, pp. 64–79.
[129] R. Nishii and S. Tanaka, "Accuracy and inaccuracy assessments in land-cover classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 37, no. 1, pp. 491–498, Jan. 1999.
[130] W. R. Tobler, "A computer movie simulating urban growth in the Detroit region," *Econom. Geogr.*, vol. 46, no. 2, pp. 234–240, Jun. 1970.
[131] M. Baatz, C. Hoffmann, and G. Willhauck, "Progressing from object-based to object-oriented image analysis," in *Object-Based Image Analysis-Spatial Concepts for Knowledge-driven Remote Sensing Applications*, T. Blaschke, S. Lang, and G. J. Hay, Eds. New York, NY, USA: Springer-Verlag, 2008, ch. 1.4, pp. 29–42.
[132] J. C. Bezdek, T. Reichherzer, G. S. Lim, and Y. Attikiouzel, "Multiple-prototype classifier design," *IEEE Trans. Syst., Man, Cybern. C, Appl. Rev.*, vol. 28, no. 1, pp. 67–79, Feb. 1998.
[133] T. Mitchell, *Machine Learning*. New York, NY, USA: McGraw-Hill, 1997.
[134] Y. V. Venkatesh and S. Kumar Raja, "On the classification of multispectral satellite images using the multilayer perceptron," *Pattern Recognit.*, vol. 36, no. 9, pp. 2161–2175, Sep. 2003.
[135] W. Kaydos, *Operational Performance Measurement: Increasing Total Productivity*. Boca Raton, FL, USA: CRC Press, 1999.
[136] M. Page-Jones, *The Practical Guide to Structured Systems Design*. Englewood Cliffs, NJ, USA: Prentice-Hall, 1988.
[137] G. Zhou and M. Kafatos, "Future Intelligent Earth Observing Satellites (FIEOS)," in *Proc. ISPRS/FIEOS Conf.*, 2002, pp. 1–8. [Online]. Available: http://www.isprs.org/proceedings/XXXIV/part1/paper/00031.pdf
[138] *ENVI 4.3, User Manual*, ITT Industries Inc., Whiteplains, NY, USA, 2006.
[139] J. Slootweg, J. P. Hettelingh, W. Tamis, and M. van't Zelfde, "Harmonizing European Land Cover Maps," Netherlands Environmental Assessment Agency, The Hague, The Netherlands, (accessed on 6 Nov. 2012). [Online]. Available: http://www.rivm.nl/bibliotheek/digitaaldepot/PBL_CCE_SR05_Chapter3.pdf
[140] O. Cerba, K. Charvat, and J. Jezek, "Data Harmonization Towards CORINE Land Cover," Univ. of West Bohemia, Pilsen, Czech Republic, (accessed on 6 Nov. 2012). [Online]. Available: www.efita.net/apps/accesbase/bindocload.asp
[141] H. Couclelis, "What geographic information science is NOT: Three theses," in *Proc. GIScience Conf.*, Columbus, OH, USA, Sep. 18–21, 2012.
[142] J. Tian and D. M. Chen, "Optimization in multi-scale segmentation of high-resolution satellite images for artificial feature recognition," *Int. J. Remote Sens.*, vol. 28, no. 20, pp. 4625–4644, Oct. 2007.
[143] L. A. Dupigny-Giroux and J. E. Lewis, "A moisture index for surface characterization over a semiarid area," *Photogramm. Eng. Remote Sens.*, vol. 65, no. 8, pp. 937–945, Aug. 1999.
[144] D. Hubel and T. Wiesel, "Receptive fields of single neurons in the cat's striate cortex," *J. Physiol.*, vol. 148, no. 3, pp. 574–591, Oct. 1959.
[145] J. Mutch and D. Lowe, "Object class recognition and localization using sparse features with limited receptive fields," *Int J. Comput. Vis.*, vol. 80, no. 1, pp. 45–57, Oct. 2008.
[146] T. N. Wiesel and D. H. Hubel, "Spatial and chromatic interactions in the lateral geniculate body of the rhesus monkey," *J. Neurophys.*, vol. 29, no. 6, pp. 1115–1156, Nov. 1966.
[147] A. Jain and G. Healey, "A multiscale representation including opponent color features for texture recognition," *IEEE Trans. Image Process.*, vol. 7, no. 1, pp. 124–128, Jan. 1998.
[148] Anonymous ftp. [Online]. Available: ftp://ftp.iluci.org/Paper/TGRS_2012_00550, otherwise, http://doi.pangaea.de/10.1594/PANGAEA.806528?format=html



Andrea Baraldi was born in Modena, Italy, in 1963. He received the Laurea (M.S.) degree in electronic engineering from the University of Bologna, Bologna, Italy, in 1989, and the Master's degree in software engineering from the University of Padova, Padova, Italy, and Purdue University, West Lafayette, IN, USA, in 1994.

From 1989 to 1990, he worked as a Research Associate at the Centro di Studio per l'Interazione Operatore-Calcolatore, National Research Council (CNR), Bologna, Italy, and served in the army at the Istituto Geografico Militare in Florence, Italy. At the European Space Agency-European Space Research Institute, Frascati, Italy, he worked as a Consultant from 1991 to 1993. From Dec. 1997 to June 1999, he was assigned with a post-doctoral fellowship in Artificial Intelligence at the International Computer Science Institute, Berkeley, CA, USA. From 2000 to 2002, as a Post-doctoral Researcher, he joined the Global Vegetation Monitoring Unit of the Institute for Environment and Sustainability (IES) of the European Commission Joint Research Center (JRC), Ispra, Italy. From 2005 to 2009, he was at the IES-Spatial Data Infrastructure, JRC. In 2009, he founded Baraldi Consultancy in Remote Sensing, a one-man company located in Modena, Italy. Since his master thesis, he has been continuing his collaboration with the Istituto di Scienze dell'Atmosfera e del Clima, CNR, Bologna, and the Istituto di Studi di Sistemi Intelligenti per l'Automazione, CNR, Bari, Italy. Currently, he is a Research Associate Professor at the Department of Geography, University of Maryland, MD, USA, involved with the automatic generation of high-level products at continental/global scale in the frame of the USGS-NASA Web-Enabled Landsat Dataset project and the NASA Land Cover and Land Use Change program. His main interests center on image understanding, with special emphasis on the development of operational, automatic, hierarchical, multiresolution, spaceborne and airborne image understanding systems consistent with human vision.

Mr. Baraldi served as an Associate Editor of the IEEE TRANSACTIONS ON NEURAL NETWORKS from 2001 to 2006.

Luigi Boschetti received the Laurea (M.S.) degree in environmental engineering and the Ph.D. degree in geodesy and geomatics from the Politecnico di Milano, Milano, Italy, in 2000 and 2005, respectively.

He was a Visiting Scientist with the Natural Resources Institute, University of Greenwich, London, U.K., from 2000 to 2002, a Research Fellow with the Institute for Environment and Sustainability, Joint Research Center, European Commission, Ispra, Italy, from 2002 to 2004, a Research Fellow with CNR-IREA, Milano, Italy, from 2004 to 2005, and a Research Scientist with the Department of Geographical Sciences, University of Maryland, College Park, MD, USA, from 2005 to 2012. He has been with the University of Idaho, Moscow, since 2012, where he is an Associate Professor with the Department of Forest, Rangeland, and Fire Sciences.

His research primarily includes the application of medium-resolution satellite data for environmental monitoring, with a focus on burned area mapping, and the development of validation techniques for global thematic products.

Michael L. Humber received the B.S. degree in geography from the University of Maryland, College Park, MD, USA, in 2011. Currently, he is working toward the M.P.S. degree in geospatial information sciences at the University of Maryland.

From 2010 to 2011, he was an Intern with the Department of Geography Research Laboratory at the University of Maryland, working on fire detection in moderate-resolution satellite imagery and global-level data visualization. Currently, he is a Graduate Research Assistant at the Department of Geography Research Laboratory at the University of Maryland, with activities focused on processing MODIS Burned Area Products, validation of maps generated from moderate to very high resolution imagery, and forest monitoring and forest cover change reporting for Central Africa.