Triple collocation to assess classification accuracy without a ground truth in case of earthquake damage assessment

Nazzareno Pierdicca, *Senior Member, IEEE*, Roberta Anniballe, Fabrizio Noto, Christian Bignami, *Member, IEEE*, Marco Chini, *Senior Member, IEEE*, Antonio Martinelli, Antonio Mannella.

***Abstract*—the assessment of satellite image classifications is usually carried out using a test sample assumed as the *ground truth*, from which the confusion matrix is derived. There are cases where the reference data, even those coming from a ground survey, are affected by errors and do not represent a reliable *truth*. In the field of geophysical parameter retrieval, the Triple Collocation (TC) technique is applied for validating remotely sensed products when the source of test data (e.g., ground data) do not represent a reliable reference. The TC is able to retrieve the error variances of three systems observing the same target parameter, assuming their errors are independent. In this paper, we exploit the same idea to test the classification accuracy in cases the *ground truth* is not available. We extend the TC approach to the classification problem for a general number of classes, but we solve it numerically for a 2-class problem (i.e., collapsed and non-collapsed buildings). The specific case refers to the detection of L'Aquila 2009 earthquake damage from Very High Resolution (VHR) optical data. The image classification, performed by exploiting an object-based analysis, is compared to two different ground surveys carried out after the earthquake by different teams and with different purposes. The work demonstrates the powerfulness of the TC approach for assessing the classification accuracy with no reliable *ground truth* available, and provides an insight into the problem of assessing damage, from satellite and on ground, in a very critical and unsafe situation, like the one occurring after an earthquake. Moreover, it was found that the remotely sensed product can have order of accuracy comparable to the ground surveys.**

***Index Terms*— classification accuracy, earthquake damage, triple colocation.**

I. Introduction

Assessing the accuracy of an image classification product is a fundamental step when remote sensing is used to produce or update thematic maps. There are many works in the literature addressing this problem [1] [2]. Speaking about thematic accuracy, a first step is the acquisition of a set of reference samples belonging to well-known classes (or categories) and the computation of the confusion matrix (CM), which counts the occurrences of each category in the classification and what is considered the *ground truth*. From the confusion matrix different quality parameters can be derived, such as the Overall Accuracy (percentage of cases correctly allocated), the Cohen's Kappa coefficient, and the user's and producer's accuracy (see [2] for a general discussion).

In many cases, getting samples of the *ground truth* is not feasible; they can be affected by errors and thus do not represent the real situation [3] [4]. Indeed, errors can affect test sets acquired from aerial images and visual interpretation used as reference to validate an automatic image classification algorithm. Moreover, even a ground survey can be affected by errors, especially when it is carried out in a difficult situation or is conceived for a purpose different from providing a reference for satellite image classification assessment. This is the case for instance of ground surveys performed after a disastrous event like an earthquake. The survey teams act in unsafe conditions (with a limited access to the affected area) with the purpose of giving a rapid assessment of the seismic intensity or checking the structural integrity of the buildings after the earthquake and their fitness for use.

In [3] this problem was deeply investigated in case of a change map, when the classification concerns the two cases of changed (Δ=1) or not changed (Δ=0) samples. The impact of imperfect reference data on the classification assessment was simulated considering errors either statistically independent or correlated to a certain degree. In order to cope with this problem, one may assume to know the quality of reference set, or to rely on latent class analysis when this information is not available at all [3]. In [5], it was shown that when the reference data are affected by errors we could still compare the accuracy of two classifications relatively if the number of test samples is large; however, the retrieved accuracies remain biased without knowing the accuracy of the reference test set.

A similar problem is faced in the field of medical diagnostic, when an accurate diagnostic test (denoted as *gold standard*) is not feasible, being too expensive or invasive. Therefore, the prevalence of a disease in a population has to be inferred by tests that have some unknown errors [6]. In particular, [7] considers a case of three tests applied to one population, a problem similar to deriving a change detection map from three different classification methods. Note that in medical diagnostic they consider generally a 2-classes problem (positive or negative test result in a patient), whereas land cover classification implies a general number of *N* categories.

In a quite different field of remote sensing, i.e., the

Manuscript received xxxxx; revised xxxx; accepted xxxxx.

N. Pierdicca and R. Anniballe are with the Department of Information Engineering, Electronics and Telecommunications, Sapienza University of Rome, Rome, Italy (e-mail: nazzareno.pierdicca@uniroma1.it).

F. Noto is with METIS s.r.l., Rome, Italy (e-mail: fabrizio.noto@gmail.com).

C. Bignami is with Istituto Nazionale di Geofisica e Vulcanologia, Rome, Italy (e-mail: christian.bignami@ingv.it).

M. Chini is with the Luxembourg Institute of Science and Technology, Belvaux, Luxembourg (e-mail: marco.chini@list.lu www.list.lu

A. Martinelli and A. Mannella are with Istituto per le Tecnologie della Costruzione of the Italian National Research Council, L'Aquila, Italy (e-mail: antonio.martinelli@itc.cnr.it, antonio.mannella@itc.cnr.it).

The work has been funded by the EC-FP7 APhoRISM project (Research, Technological Development and Demonstration Activities, grant agreement n. 606738). We thank the Italian Department of Civil Protection for providing the ground survey.

Digital Object Identifier: ??????

geophysical parameter retrieval, the Triple Collocation (TC) technique is successfully applied for validating remotely sensed products when the true values of the target parameter are not surely known, as the source of validation data is known to have its own errors. The TC is able to retrieve the error variances of three systems observing the same target parameter, assuming their errors are independent. The technique has been originally conceived to assess the accuracy of wind speed retrieval over sea [8], and successively largely adopted for testing soil moisture retrieval from space sensors or hydrological models [9], [10], even combining retrievals from four systems [11].

In this paper, we exploit the TC idea to test the classification accuracy in case the *ground truth* is not available. In this case, the TC approach consists in comparing three different classifications of the same test samples to infer their CMs with respect to the unknown class the samples actually belong to.

The proposed novel TC for Classification Assessment (TCCA) method represents a manageable way to solve a latent class model [3]. We formulate it for a general number N of classes, and then solve it numerically for the 2-class problem. In particular, we consider the problem of detecting the building collapses from Very High Resolution (VHR) optical images collected before and after the destructive earthquake that hit the city of L'Aquila, Italy, in 2009. The damage classification map was compared to two different ground surveys carried out after the earthquake by different Institutions and with different purposes and protocols.

Section II presents the TCCA approach, with all the mathematical details included in the Appendices. Section III describes the data, both the ground surveys and the satellite images and their classification algorithms. Section IV describes the results of the TCCA, and Section V draws the final conclusions.

II. The Triple Colocation Approach

TC is used for validating geophysical parameter retrievals (e.g., wind speed, soil moisture) assuming that three systems (X, Y, Z), measuring the same target parameter θ, are affected by systematic calibration and additive random errors. Let us consider one system as the reference (i.e., unitary gain and zero bias) and assume the additive random errors are statistically independent and independent from the true parameter θ. Three variance-covariance matrices can be computed from the three sets of observations x, y, z of the same target parameter. It can be demonstrated that the unknown gains and variances of the random errors affecting each system can be derived. Although slight differences in the solution may exist, as discussed in [10]. the fundamental hypothesis of uncorrelated errors is common to most papers on this topic.

Considering the problem of image classification, each X, Y, and Z classification system associates to each sample a label x, y, z, respectively, which is an integer between 1 and N, for an N-classes problem. Since we do not have a *gold standard*, all of them are imperfect indicators of the unobserved (latent) status θ of the samples [3]. We can therefore build three different CM's from each pair of classification results. This can be translated into a probabilistic formulation as follows. If we normalize the three CM's by the number of samples in the test set (getting the three normalized confusion matrices NCM's, from now on denoted as XY, XZ and YZ), we get an estimate of the join probabilities $P_{X,Y}(x,y)$, $P_{X,Z}(x,z)$, $P_{Y,Z}(y,z)$. With regard to XY, and similarly for the others, one can write:

$$P_{XY}(x,y) \leftrightarrow \mathrm{XY} = \begin{bmatrix} p_{11}^{XY} & p_{12}^{XY} & \cdots & p_{1N}^{XY} \\ p_{21}^{XY} & p_{21}^{XY} & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots \\ p_{N1}^{XY} & p_{N2}^{XY} & \cdots & p_{NN}^{XY} \end{bmatrix} \quad (1)$$

Then, XY contains the conditional probabilities arranged in a matricial form.

Actually, one wants to know the joint probabilities $P_{X,\Theta}(x,\theta)$, $P_{Y,\Theta}(y,\theta)$, $P_{Z,\Theta}(z,\theta)$ of the three system outcomes with respect to the latent variable, i.e., the NCM with respect to the class the samples really belong to. With reference to system X, we should retrieve XΘ, which corresponds to the joint probability $P_{X,\Theta}(x,\theta)$:

$$P_{X\Theta}(x,\theta) \leftrightarrow \mathrm{X\Theta} = \begin{bmatrix} p_{11}^{X} & p_{12}^{X} & \cdots & p_{1N}^{X} \\ p_{21}^{X} & p_{22}^{X} & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots \\ p_{N1}^{X} & p_{N2}^{X} & \cdots & p_{NN}^{X} \end{bmatrix} \quad (2)$$

where rows indicate the outcomes of the classifier and columns the true classes, so that p_{ij}^{X} is the joint probability of getting class i out of the classifier, being j the true class. The diagonal terms indicate the probability of correct classification. Similarly, we refer to YΘ and ZΘ as the NCM of the other systems with elements p_{ij}^{Y} and p_{ij}^{Z}.

By adapting the hypothesis of TC, we assume that the errors of each classification system are conditionally independent to the errors of the other systems (e.g., the outcome of one classifier cannot be deduced from the outcome of the others) (see also [3] for a discussion on this hypothesis). It is demonstrated in the Appendix-A that we can derive the following equations for each target NCMs XΘ, YΘ and ZΘ we want to retrieve, which are expressed as function of the known matrices XY, XZ, YZ derived from the test set:

$$\begin{aligned} \mathrm{X\Theta} \cdot \mathrm{P} \cdot \mathrm{X\Theta}^{\mathrm{T}} &= \mathrm{XZ} \cdot \mathrm{YZ}^{-1} \cdot \mathrm{XY}^{\mathrm{T}} \\ \mathrm{Y\Theta} \cdot \mathrm{P} \cdot \mathrm{Y\Theta}^{\mathrm{T}} &= \mathrm{YZ} \cdot \mathrm{XZ}^{-1} \cdot \mathrm{XY} \\ \mathrm{Z\Theta} \cdot \mathrm{P} \cdot \mathrm{Z\Theta}^{\mathrm{T}} &= \mathrm{YZ}^{\mathrm{T}} \cdot \mathrm{XY}^{-1} \cdot \mathrm{XZ} \end{aligned} \quad (3)$$

where superscripts "T" and "-1" indicate transposition and matrix inversion, respectively. Matrix P is a diagonal matrix containing the inverse of N unknown probabilities p_j (j=1, … N) of the classes (sometime denoted as prevalence of the classes), which can be computed from any of the target NCMs, so that the following 3N constraints also apply:

$$p_j = \sum_{i=1}^{N} p_{ij}^{X} = \sum_{i=1}^{N} p_{ij}^{Y} = \sum_{i=1}^{N} p_{ij}^{Z} \quad (4)$$

Note that matrices in equations (3) (e.g., $\mathrm{X\Theta} \cdot \mathrm{P} \cdot \mathrm{X\Theta}^{\mathrm{T}}$) are symmetric, so each matricial equation corresponds to $N(N+1)/2$ polynomial equations in the N^2 unknown $p_{ij}^{X,Y,Z}$, plus 3 constraints requiring that by saturating the joint probability we get one, i.e., $\sum\sum p_{ij}^{X} = \sum\sum p_{ij}^{Y} = \sum\sum p_{ij}^{Z} = 1$.

In TTCA we can also build a 3-dimensional NCM, denoted

as XYZ, representing the joint probability $P_{X,Y,Z}(x,y,z)$, with elements p_{ijk} being the probability of encountering a sample labeled as classes i, j, and k by the three classifiers, respectively. This leads to other constraints for the target NCM's (see Appendix-A):

$$p_{i,j,k} = \sum_{m=1}^{N} \frac{p_{im}^{X} p_{jm}^{Y} p_{km}^{Z}}{p_m^2} \quad (5)$$

which are not all independent, as by saturating with respect to the classes one obtains $\sum_k p_{i,j,k} = p_{i,j}^{XY}$, $\sum_j p_{i,j,k} = p_{i,j}^{XZ}$, $\sum_i p_{i,j,k} = p_{i,j}^{YZ}$.

If we expand the matricial relations (3) we get, together with eq. (4) and (5) a polynomial system of equations in the unknown probabilities $p_{ij}^{X,Y,Z}$. The analysis of its solution (i.e., existence, unicity) for the general case of N classes is beyond the scope of this paper and is left to further work. Here, we face the case of a 2-class problem, which is typical of any change detection problem, e.g., in case of collapsed building detection after an earthquake. The solution is found in Appendix-B. It requires to solve the equation (B.5) for the prevalence of class i=1 (i.e., p_1), where $p_{11}^{X,Y,Z}$, $p_{12}^{X,Y,Z}$, $p_{21}^{X,Y,Z}$, $p_{22}^{X,Y,Z}$ as function of p_1 are given by equations (B.4). We will use this method in section IV in the case of L'Aquila damage classifications provided by satellite remote sensing and two ground surveys.

The hypothesis of conditional independence of the errors of the three systems is often assumed when exploiting the latent class analysis, as discussed in [3]. However, in that paper it was also investigated the presence of correlated errors and concluded that the satisfaction of the model assumption is critical in its use. Solutions to this issue were proposed for TC, for instance using four systems instead of three [12]. This issue for TCCA is left for future studies.

III. The Earthquake Damage Detection Case Study

A. Data

On April 6, 2009 at 1:32 GMT, an earthquake hit L'Aquila city, in Central Italy. The main shock was rated 6.3 on the moment magnitude (Mw) scale. Two ground surveys were carried out after the event and made available for this study. The first one is the survey performed by Istituto Nazionale di Geofisica e Vulcanologia (INGV) Macroiseismic team (QUEST - QUick Earthquake Survey Team, see http://quest.ingv.it), while the second one was carried out by the Italian Department of Civil Protection (DPC).

INGV researchers, in one week of fieldwork, collected information on type of buildings and the suffered damage, according to the European Macroseismic Scale 1998 (EMS'98) [13]. The damage grade ranges from 0 to 5, i.e., from no damage to completely collapsed, and it was attached as an attribute to a Geographical Information System (GIS) layer, with polygons representing the building footprint. More than 1600 buildings were surveyed in the central part of the town, and 74 of them were found damaged with grade 5. It is worth noticing that the data were collected by a visual inspection, building-by-building, only looking from outside the buildings itself, because INGV teams were not allowed to go inside the edifices for safety reasons. The purpose of INGV survey was the estimation of the Seismic Intensity of the earthquake.

The DPC survey was carried out during the six months following the seismic event and includes a detailed review of the interior parts of the buildings. Differently from INGV, the final goal of this survey was to classify the building usability and assess the post-earthquake damage. A different survey protocol was followed for private and public edifices. The form ("Agibilità e Danno sugli Edifici pubblici e privati": AeDES) filled in for each private building, contains more than 250 fields describing the structural typology and the damage grade for each structural component. A different form was filled for the settlements classified as cultural heritage. From this plenty of information a damage indicator and a vulnerability class were calculated following the EMS '98 scale by Istituto per le Tecnologie della Costruzione (ITC) of the Italian National Research Council (CNR) and attached to a GIS map of the town. Considering the central part of the town the DPC survey consists of 2003 buildings, of which 129 were associated to damage grade 5.

Figure 1 and figure 2 show the two surveys, superimposed to a satellite image, by colour coding each building according to its EMS'98 damage grade.

Very High Resolution (VHR) optical images were collected by the QuickBird satellite before (September 4, 2006), and 2 days after (April 8, 2009) the catastrophic event. Each acquisition consists of a panchromatic (PAN) and a multispectral (MS) image (blue, green, red and near-infrared channels). Nominally, at nadir, the spatial resolution of the PAN data is 0.6 m, while the MS image has a 2.44 m resolution.

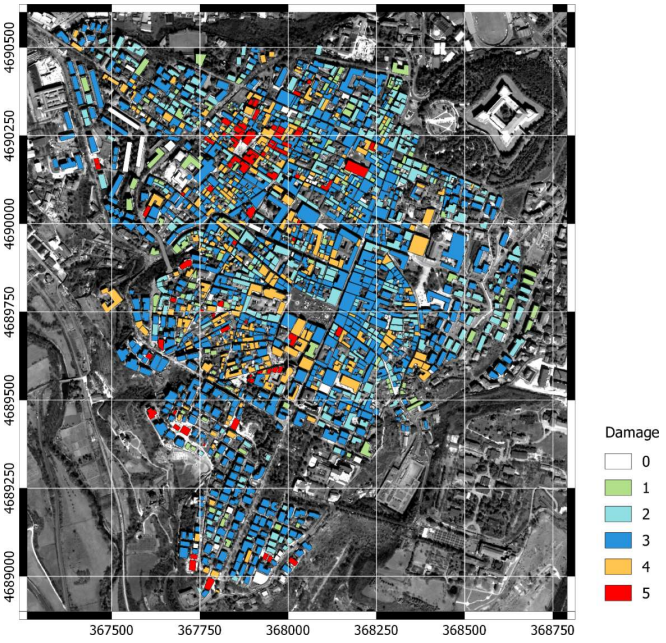

Figure 1. Damage distribution of L'Aquila city center according to INGV survey. The polygons of surveyed buildings are superimposed on a very high resolution panchromatic image acquired by the QuickBird satellite.

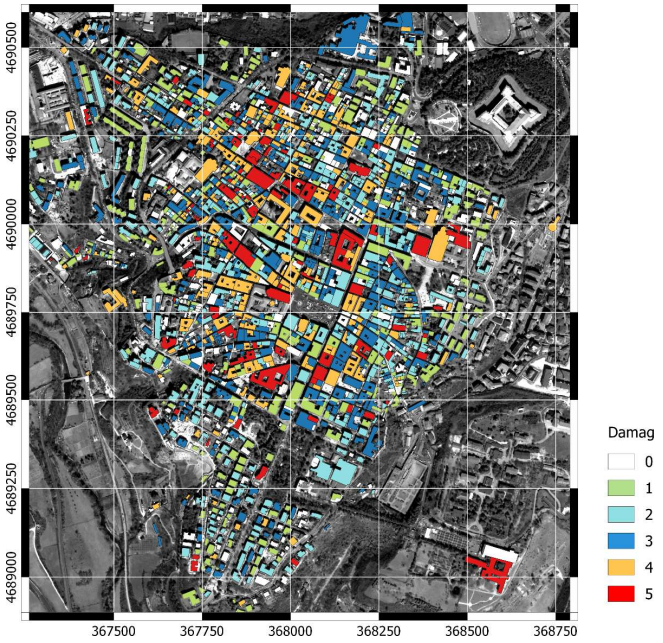

Figure 2. As in Figure 1, but showing the DPC survey.

B. *Image processing approach*

To classify the QuickBird images we adopted an object based analysis approach. The available GIS layer of polygons representing the building footprints was used to segment the image and identify the pixels belonging to each building. A precise registration of the two images and orthorectification with respect to the GIS layer was carried out using a Digital Terrain Model (DTM) provided by a LiDAR scanner survey of the area. For each building (i.e., an image object) a number of features were computed from the pre- and post-seismic data, both panchromatic and multispectral, potentially capable of detecting changes. Several change detection features have been tested and the six most sensitive to changes caused by the earthquake have been considered in the final classification procedure:

- Mutual Information (MI) [14] between the two panchromatic images: it provides a measure of the pixel intensity correlation per objects across the two days;
- Textural parameters difference between pre-seismic and post-seismic images:Energy, Correlation, Homogeneity, Contrast have been considered and computed within each object with 1 pixel shift [15], representing changes in the spatial arrangement of pixel intensities;
- Difference of Saturation color attribute derived from multispectral images [16]: it helps detecting changes not associated to damage, for instance in case of building restoration.

Note that we considered a set of 6 change features for each object. All those features were processed by a supervised classification algorithm considering two classes, corresponding to *collapsed* buildings (grade D=5 in the EMS ‘98 scale) and *non-collapsed* ones (grade D< 5 in the EMS ‘98 scale). The training set was extracted by a visual inspection of the Quickbird image itself.

Let $\Omega = \{\omega_1, \omega_2, ...\omega_d\}$ be the change parameter vector (with d=6) extracted from the pre- and post-event optical images and associated to a building. Let $\Theta = \{\theta_1, \theta_2\}$ be the set of classes, i.e., D=5 or D<5. Following the Maximum Likelihood (ML) criteria, given the observed feature vector Ω, we assign the building to the class θ_k (k=1,2) with the highest class-conditional probability density function (pdf) (also called class likelihood function) $P(\Omega|\theta_k)$.

As for the class-conditional pdf’s, they are estimated from a training set trough the non-parametric approach known as Parzen window method, assuming, moreover, the class-conditional independence of all the features, i.e., the Naïve hypothesis [17]. In detail, given a set of n_k training samples from class k, and assuming a Gaussian kernel, the Parzen window method estimates the distribution $P(\Omega|\theta_k)$ by the following:

$$p(\Omega|\theta_k) = \prod_{i=1}^{d} \frac{1}{n_k h_i} \sum_{j=1}^{n_k} \exp\left(-\frac{\left(\omega_i - \omega_{ij}^{(k)}\right)^2}{2h_i} \right) \qquad (6)$$

where $\omega_{ij}^{(k)}$ is the jth observation of the ith feature from the kth class, and h is the so called bandwidth parameter. Equation (6) shows that, for each feature, the pdf of the data given the class is estimated as sum of Gaussian kernel functions placed on each training data point.

IV. TCCA APPLIED TO THE L’AQUILA TEST CASE

We applied the TCCA method to the three damage maps from DPC, INGV and from Quickbird image classification assuming the three systems provide independent classification maps. Note that the fact that the two surveys were carried out by different teams and the use of an independent supervised image processing approach guarantee that the hypothesis of independent errors of the systems holds.

Since the two ground surveys do not refer exactly to the same building GIS layer, in order to apply the TCCA approach it was first necessary to intersect the two layers; this leaded to a common set of 1445 buildings with associated labels, i.e., *collapsed* (D=5) or *non-collapsed* (D<5), according to the two surveys. The ML classification was then applied to the same set of image objects.

The CM’s resulting from the comparison of the ML algorithm result and the classification provided by the ground surveys are reported in Table I (2-D *standard* CM’s), and Table II (3-D CM providing all combinations of classification outcomes). Table I shows the standard confusion matrix considering all combinations of two classifications (main diagonal counts number of objects with same classification outcome, whilst off-diagonal terms are number of mismatches). Table II refers to the triple collocation approach, as it counts the occurrences of all possible combinations of the outcomes of the three classifiers. Specifically, it reports the number of samples labeled as D=5 by the three classifiers, or commonly labeled as D<5, or labeled as D=5 by two classifiers and D<5 by the third one, and so on so forth.

TABLE I
2×2 CONFUSION MATRICES FOR THE THREE PAIRS OF SYSTEMS, ASSUMING X AS THE DPC SURVEY, Y THE INGV SURVEY AND Z THE EARTH OBSERVATION IMAGE CLASSIFICATION (EO).

DPC \ INGV	*D*<5	*D*=5
D<5	1309	30
D=5	75	31

DPC \ EO	*D*<5	*D*=5
D<5	1297	42
D=5	81	25

INGV \ EO	*D*<5	*D*=5
D<5	1341	43
D=5	37	24

TABLE II
3-D CONFUSION MATRIX FOR THE THREE SYSTEMS AS DEFINED IN TABLE I

	INGV D<5		INGV D=5	
	EO D<5	EO D=5	EO D<5	EO D=5
DPC D<5	1276	33	21	9
DPC D=5	65	10	16	15

We can notice that the performances of the image classification (denoted as EO in the tables) compared to the two ground surveys, supposing they could be considered as the *ground truth*, are not very good. Only 24 collapses were detected out of 61 according to INGV, or 25 out of 106 according to DPC, with overall accuracy of 94.5 % (Cohen's Kappa, K=0.33) and 91.5 % (K=0.18), respectively. However, the matching between the two ground surveys, which we remind were carried out according to different purposes, was not good as well, being 92.7 % the percentage of buildings with the same classification outcome (K=0.36). Then the question we want to answer by using the TCCA is: which is the actual accuracy of the three systems with respect to the *truth*?

Using the TCCA solution for the 2-class problem, after normalization of the CM's in Table I, we derived the three NCM's XΘ, YΘ, ZΘ as function of the probability of collapsed building p_2 using eq. (B.4). Then from the NCM's we computed the Overall Accuracy and Cohen's Kappa coefficients [1][2]. In Fig. 3 it is shown that the INGV classification has a Kappa coefficient always greater than the others, despite of the true prevalence of *damaged*, class. Concerning the other two classification results EO is slightly better than the DPC classification when the probability of damage is low, whilst DPC becomes a bit better respect the EO classification if the probability of damage is higher (see Fig. 3). This result can be justified if one considers that according to the CM's in Table I the INGV classification has a fair matching with both DPC and EO (92.7% and 94.5%, respectively), whereas this is not true for the others that exhibit a pronounced disagreement (91.5% between EO and DPC). This experimental evidence makes the TCCA solution trust more on the INGV classification, as it can be intuitively understood. If the prevalence of *damaged* class were high, the TCCA would have trusted more on the DPC survey that has a higher occurrence of D=5 class respect to EO.

Fig. 4 compares the Overall Accuracy of the systems. It is noticeable that according to this quality score the satellite classification has a quality intermediate when compared to the ground surveys.

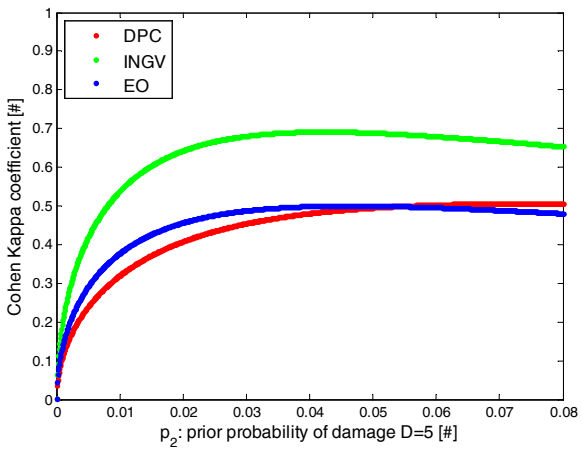

Figure 3. Cohen Kappa coefficient of systems DPC, INGV and image classification results (EO) as function of prior probability of collapsed building.

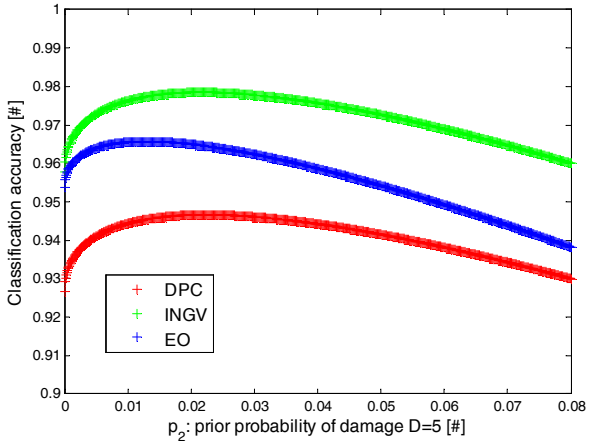

Figure 4. Overall Accuracy of systems DPC, INGV and image classification results (EO) as function of prior probability of collapsed building.

To retrieve the final solution the probability of collapse p_2 has to be derived by solving eq. (B.5). From the CM in Table II it comes out that the number of samples classified as *undamaged* by the three systems is 1276, whereas the number of samples accordantly considered as *damaged* is 15. Then, dividing by 1445, it comes out that p_{111}= 0.8830 and p_{222}=0.0104. A numerical solution of eq. (B.5) was found by using the Matlab© Symbolic Tool. Here for sake of better clarity we depicts a graphical solution in Fig. 5, where $p_{21}^X p_{21}^Y p_{21}^Z / p_1^2 + p_{22}^X p_{22}^Y p_{22}^Z / (1-p_1)^2$ derived by using (B.4) is plotted as function of p_2, so that according to (B.5) the solution is found when the value $p_{222} = 0.0104$ is reached. Both methods provided p_2=0.0529, from which XΘ, YΘ and ZΘ and associated classification accuracy scores were finally retrieved using the eq. (B.4).

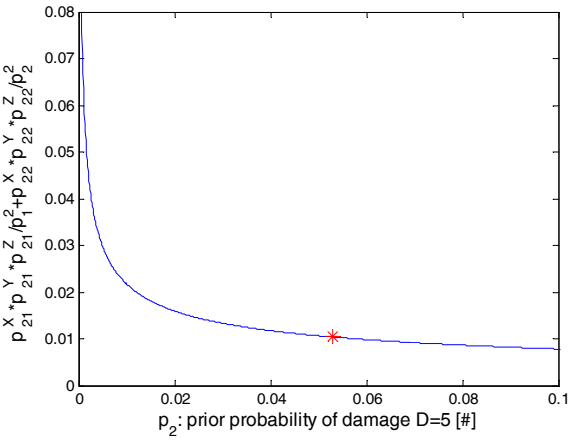

Figure 5. $p_{222} = p_{21}^X p_{21}^Y p_{21}^Z / p_1^2 + p_{22}^X p_{22}^Y p_{22}^Z / (1-p_1)^2$ as function of p_2. Red dot corresponds to the solution of eq. (B.5) with p_{222}=15/1445=0.0104.

Multiplying XΘ, YΘ and ZΘ by the number of samples, we computed the CM's with respect to the *truth*, which are reported in Table III. They provide the number of correct classifications, misdetections ad false alarms with respect to the unobserved true status of the samples. Note that column totals are the estimated number of *damaged* (D=5) and *non-damaged* (D<5) buildings. From those CM's we can derive the Overall Accuracy scores and the Cohen's Kappa coefficients of the system.

TABLE III
CONFUSION MATRICES OF THE THREE DAMAGE CLASSIFICATIONS WITH RESPECT TO THE "TRUE". X REFERS TO THE DPC SURVEY, Y TO THE INGV SURVEY AND Z TO THE IMAGE CLASSIFICATION (EO). THEY REPRESENT THE NUMBER OF CORRECT CLASSIFICATIONS, MISDETECTION AND FALSE ALARMS WITH REFERENCE TO THE UNOBSERVED TRUE STATUS OF THE SAMPLES. COLUMN TOTALS IS THE ESTIMATED NUMBER OF REALLY *DAMAGED* (D=5) AND *NON-DAMAGED* (D<5) BUILDINGS.

DPC	Truth	
	D<5	*D=5*
D<5	*1311*	*28*
D=5	58	48

INGV	Truth	
	D<5	*D=5*
D<5	*1356*	*28*
D=5	13	48

EO	Truth	
	D<5	*D=5*
D<5	*1339*	*38*
D=5	30	38

We observe again that the INGV survey is the most accurate (97.2 % accuracy, 0.687 Cohen's Kappa) with few false positives but a bit more misdetections of collapse. The image classification has worse performances (accuracy of 95.3 %, Cohen's Kappa of 0.500), with a much larger number of false alarms. However, its real accuracy is better than that found by comparing directly with the two ground surveys. The DPC survey, with an accuracy of 94.1 % and a Cohen's Kappa of 0.499, is a bit worse. This result can be justified by the different purposes of the surveys. In particular, the DPC survey provides a great amount of information which aims at assessing the structural condition of the buildings and their fitness for use, despite of their exterior appearance. For instance, a heavy damage of the vertical structures (e.g., pillars) may have been considered as a predominant damage of grade 5, even if not consistent with the general definition of EMS'98 when looking at the building from the exterior. Conversely, the image based classification suffers from situations we were able to identify in some cases by a careful analysis of the satellite images but also from an aerial survey and ground based photographs. Heavy damage may be not visible looking down upon, for instance in case of the so-called *pancake* effect, which we recognized occurred in some buildings, and is characterized by the downfall of the ground floor only. In Fig. 6 a picture of a building taken from ground and the same building imaged by Quickbird after the earthquake are shown to give an example of this effect. Additionally, false positives occurred when the building underwent a restoration just after the acquisition of the pre-event image, which was wrongly classified as a change due to the earthquake. Despite of these problems and the difficulty of change detection in a very dense historical town, the satellite classification exhibited a damage detection performance comparable or even better of at least one ground survey.

Figure 6. Photograph of a building in L'Aquila affected by the *pancake* type of damage and the same building imaged by the post-event Quickbird data take. With respect to the pre –

event image (not shown), the only change is the presence of dust on the street due to the damage that was classified as EMSL'98 grade *D*=5.

V. Conclusions

We have derived a triple colocation approach (TCCA) to retrieve the accuracy of three classifications in the absence of a *ground truth* (or when any reference data or ground surveys are known to be affected by errors). Once derived the confusion matrices among each pair of classifications, with the hypothesis of independent errors, it is possible to write a set of polynomial equations for the confusion matrices with respect to an unknown truth.

The method was applied to the classification of urban damage after the L'Aquila 2009 earthquake, based on the EMS'98, comparing two independent ground surveys and an automatic supervised classification of satellite VHR Quickbird images. Although one can expect a ground survey would represent a reference for validating satellite image classifications, the TCCA investigation showed that assuming a ground survey as the reference for testing the satellite damage classification can be questionable. The concept of damaged or non-damaged can be controversial on its own, as it may be related to the purpose of the survey and to the conditions in which the teams were operating. The classification from satellite images provided accuracy in distinguishing EMS'96 damage grade 5 comparable to at least one of the surveys. Indeed, the satellite has potentially a more rapid response and undoubtedly much less cost of a ground survey, so that it may play a significant role in disaster managing.

When evaluating the retrieved classification performances in this case, it must be considered the challenges of mapping damage in the dense settlement of L'Aquila historical town. In any case, this work paves the way to a more careful assessment of earthquake damage maps from remote sensing data.

Appendix A: TCCA Analytical Formulation

The hypothesis of independent errors requires that the probability of outcome *x, y, z* of each system is not dependent on the result of the others, e.g., $P_X(x|y,\theta) = P_X(x|\theta)$. Then by using the Bayesian theorem we can write for the pair of systems X and Y (in the sequel we omit the subscript in the pdf symbol *P*() for sake of brevity):

$$P(x,y,\theta) = P(x|y,\theta)P(y,\theta) = P(x|\theta)P(y,\theta) = \frac{P(x,\theta)P(y,\theta)}{P(\theta)} \quad \text{(A.1)}$$

By saturating with respect to θ we get the joint probability *P*(*x*,*y*), i.e., the observed confusion matrix XY whose elements can therefore be expressed as:

$$p_{i,j}^{XY} = \sum_{k=1}^{N} \frac{p_{ik}^{X} p_{jk}^{Y}}{p_k} \quad \text{(A.2)}$$

Similarly for the other pairs. It is straightforward to verify that this corresponds to write the following three matricial relations relating the target normalized confusion matrices XΘ, YΘ and ZΘ we want to retrieve to the observed ones, i.e., XY, XZ, YZ:

$$\begin{aligned} \mathrm{X\Theta} \cdot \mathrm{P} \cdot \mathrm{Y\Theta}^{\mathrm{T}} &= \mathrm{XY} \\ \mathrm{X\Theta} \cdot \mathrm{P} \cdot \mathrm{Z\Theta}^{\mathrm{T}} &= \mathrm{XZ} \\ \mathrm{Y\Theta} \cdot \mathrm{P} \cdot \mathrm{Z\Theta}^{\mathrm{T}} &= \mathrm{YZ} \end{aligned} \quad \text{(A.3)}$$

P is a N×N diagonal matrix given by:

$$\mathrm{P} = \begin{bmatrix} 1/p_1 & 0 & \dots & 0 \\ 0 & 1/p_2 & \dots & \dots \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & 1/p_N \end{bmatrix} \quad \text{(A.4)}$$

Where p_j (j=1, … N) are the N unknown probabilities of the classes. By deriving XΘ from the second equation in (A.3) and YΘ from the third equation, substituting into the first equation and using well-known matrix operation rules, one obtains:

$$\mathrm{X\Theta} \cdot \mathrm{P} \cdot \mathrm{Y\Theta}^{\mathrm{T}} = \mathrm{XZ} \cdot [\mathrm{P} \cdot \mathrm{Z\Theta}^{\mathrm{T}}]^{-1} \cdot \mathrm{P} \cdot [\mathrm{YZ} \cdot [\mathrm{P} \cdot \mathrm{Z\Theta}^{\mathrm{T}}]^{-1}]^{T} = \mathrm{XY}$$

$$\mathrm{XZ} \cdot [\mathrm{Z\Theta}^{\mathrm{T}}]^{-1} \cdot \mathrm{P}^{-1} \cdot \mathrm{P} \cdot [\mathrm{YZ} \cdot [\mathrm{Z\Theta}^{\mathrm{T}}]^{-1} \mathrm{P}^{-1}]^{T} = \mathrm{XY}$$

$$\mathrm{XZ} \cdot [\mathrm{Z\Theta}^{\mathrm{T}}]^{-1} \cdot \mathrm{I} \cdot \mathrm{P}^{-1} \cdot \mathrm{Z\Theta}^{-1} \cdot \mathrm{YZ}^{\mathrm{T}} = \mathrm{XY}$$

Where I is the identity matrix and we considered that P is diagonal so that $\mathrm{P}^{-1} \cdot \mathrm{P} = \mathrm{I}$. Then multiplying by XZ^{-1} on the left and by $[\mathrm{YZ}^{\mathrm{T}}]^{-1}$ on the right:

$$[\mathrm{Z\Theta}^{\mathrm{T}}]^{-1} \cdot \mathrm{P}^{-1} \cdot \mathrm{Z\Theta}^{-1} = \mathrm{XZ}^{-1} \cdot \mathrm{XY} \cdot [\mathrm{YZ}^{\mathrm{T}}]^{-1}$$

$$\mathrm{Z\Theta} \cdot \mathrm{P} \cdot \mathrm{Z\Theta}^{\mathrm{T}} = \mathrm{YZ}^{\mathrm{T}} \cdot \mathrm{XY}^{-1} \cdot \mathrm{XZ} \quad \text{(A.5)}$$

Which is the third of eq. (3) in the main text we wanted to demonstrate.

Similarly for the joint probability $P(x,y,z,\theta)$ and using (A.1) one can write:

$$P(x,y,z,\theta) = P(x,y|z,\theta)P(z,\theta) = P(x,y|\theta)P(z,\theta) = \frac{P(x,y,\theta)P(z,\theta)}{P(\theta)} = \frac{P(x,\theta)P(y,\theta)P(z,\theta)}{P^2(\theta)} \quad \text{(A.6)}$$

Which leads to eq. (5) in the main text once we saturate with respect to θ.

Appendix B: TCCA Solution for *N*=2 classes

Eqs. (3), (4) and (5) can be easily solved in case of N=2 classes. We show the solution for XΘ only, considering it is identical for the other CMs.

Each right term in eq. (3) is a symmetric 2×2 known matrix (it is a function of the 3 CM's derived from the test set), whose elements are hereafter denoted as *a, b, c*. Then, eq. (3) becomes (we omits superscript "X" for sake of simplicity):

$$\begin{bmatrix} p_{11} & p_{12} \\ p_{21} & p_{22} \end{bmatrix} \begin{bmatrix} \frac{1}{p_1} & 0 \\ 0 & \frac{1}{p_2} \end{bmatrix} \begin{bmatrix} p_{11} & p_{21} \\ p_{12} & p_{22} \end{bmatrix} = \begin{bmatrix} a & b \\ b & c \end{bmatrix} \quad \text{(B.1)}$$

Then, using eq. (4) the probabilities of each class are $p_1 = p_{11} + p_{21}$, $p_2 = 1 - p_1 = p_{12} + p_{22}$. Eq. (B.1) corresponds to three scalar equations, which are dependent since from (B1) it follows:

$$\begin{bmatrix} \frac{p_{11}^2}{p_1}+\frac{p_{12}^2}{1-p_1} & \frac{p_{11}p_{21}}{p_1}+\frac{p_{12}p_{22}}{1-p_1} \\ \frac{p_{11}p_{21}}{p_1}+\frac{p_{12}p_{22}}{1-p_1} & \frac{p_{21}^2}{p_1}+\frac{p_{22}^2}{1-p_1} \end{bmatrix} = \begin{bmatrix} a & b \\ b & c \end{bmatrix}$$

Indeed, it can be easily found that $a+b=p_{11}+p_{12}$, $b+c=p_{21}+p_{22}$, and then $a+2b+c=1$. Therefore, we solve the equation with respect to an unknown quantity, that we assume being p_1. As for the upper-left element of (B.1) one can write:

$$\frac{p_{11}^2}{p_1}+\frac{p_{12}^2}{1-p_1}=a$$

Then:

$$ap_1(1-p_1)=(1-p_1)p_{11}^2+p_1p_{12}^2 \qquad \text{(B.2)}$$

Using $p_{12}=a+b-p_{11}$ and $c=1-a-2b$, we found a second-degree algebraic equation in p_{11}, whose coefficients are functions of p_1:

$$p_{11}^2-2p_1(a+b)p_{11}+p_1(ap_1+b^2-ac)=0 \qquad \text{(B.3)}$$

That we can easily solve, and then we can derive all the elements of XΘ:

$$\begin{gathered} p_{11}=p_1(a+b)\pm\sqrt{p_1^2(a+b)^2-p_1(ap_1+b^2-ac)} \\ p_{12}=a+b-p_{11} \\ p_{21}=p_1-p_{11} \\ p_{22}=b+c-p_{21} \end{gathered} \qquad \text{(B.4)}$$

The sign shall be chosen considering the constraint of p_{ij} being real and between zero and one.

To found p_1 to be substituted into eq. (B.4) we have to rely on the 2-D CM and to eq. (5). It is enough to impose only one of the eight constraints of eq. (5) since they are all dependent, so for instance we can consider:

$$p_{222}=\frac{p_{21}^X p_{21}^Y p_{21}^Z}{p_1^2}+\frac{p_{22}^X p_{22}^Y p_{22}^Z}{(1-p_1)^2} \qquad \text{(B.5)}$$

which can be easily solved numerically, as done in the main text.

REFERENCES

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

[2] Foody, G.M., "Status of land cover classification accuracy assessment", *Remote Sensing of Environment*, Vol. 80, Issue 1, pp. 185–201, April 2002

[3] Foody, G. M., "Assessing the accuracy of land cover change with imperfect ground reference data", *Remote Sensing of Environment*, Vol. 114, Issue 10, pp. 2271–2285, October 2010.

[4] Baraldi A., L. Bruzzone, P. Blonda, "Quality Assessment of Classification and Cluster Maps Without Ground Truth Knowledge", *IEEE Trans. Geosc. Rem. Sens*., Vol. 43, N. 4, pp. 857-873, April 2005.

[5] Carlotto M. J., "Effect of errors in ground truth on classification accuracy", *International Journal of Remote Sensing*, Vol. 30, n. 18, pp. 4831-4849, 2009.

[6] Rutjes A.W.S., Reitsma J.B., Coomarasamy A., Khan K.S. and Bossuyt P.M.M., "Evaluation of diagnostic tests when there is no gold standard. A review of methods", *Health Technology Assessment*, Vol. II, No. 50, 2007.

[7] Sullivan Pepe M., Holly J., "Insights into latent class analysis of diagnostic test Performance", Biostatistics, 8, 2, pp. 474–484, 2007.

[8] A. Stoffelen, "Toward the true near-surface wind speed: Error modelling and calibration using triple collocation", *J. Geophys. Res*., 103, 7755–7766, 1998.

[9] W. A. Dorigo , K. Scipal, R.M. Parinussa, Y.Y. Liu, W. Wagner, R. A. M. de Jeu, V. Naeimi, "Error characterization of global active and passive microwave soil moisture datasets", *Hydrol. Earth Syst. Sci*, vol. 14, pp. 2605-2616, 2010.

[10] N. Pierdicca, F. Fascetti, L. Pulvirenti, R. Crapolicchio, J. Muñoz-Sabater, "Analysis of Ascat, SMOS, In-situ and Land Model soil moisture as a regionalized variable over Europe and North Africa", Remote Sensing of Environment, *Remote Sensing of Environment*, 170, 28, 2015a.

[11] Pierdicca N., F. Fascetti, L. Pulvirenti, R. Crapolicchio, J. Munoz-Sabater, "Quadruple Collocation Analysis for Soil Moisture Product Assessment", *IEEE Geosc.and Rem. Sens. Lett.*, Vol 12, pp. 1595-1599, 2015b.

[12] Gruber, A., Su, C.-H., Crow, W. T., Zwieback, S., Dorigo, W. A., & Wagner, W., "Estimating error cross-correlations in soil moisture data sets using extended collocation analysis", *Journal of Geophysical Research: Atmospheres*, 121(3), 1208–1219, 2016.

[13] Grünthal G. (ed.), "European Macroseismic Scale 1998 (EMS-98)", Cahiers du Centre Européen de Géodynamique et de Séismologie", Vol. 15, 99 pp., 1998.

[14] Erten, E., Reigber, A., Ferro-Famil, L., & Hellwich, O., "A new coherent similarity measure for temporal multichannel scene characterization", *IEEE Transactions Geosc. Rem. Sens*., Vol. 50, N.7, pp. 2839-2851, 2012.

[15] Haralick, R. M., Shanmugam, K., & Dinstein, I. H., "Textural features for image classification", , *IEEE Trans. Systems, Man and Cybernetics*, Vol. 6, pp. 610-621, 1973

[16] Smith, A. R., "Color gamut transform pairs", *ACM Siggraph Computer Graphics*, Vol. 12. No. 3, 1978.

[17] Parzen, E., "On estimation of a probability density function and mode. The annals of mathematical statistics", 33(3), 1065-1076, 1962.