Capacity and limits of multimodal remote sensing: theoretical aspects and automatic information theory-based image selection

Saloua Chlaily, *Member, IEEE,* Mauro Dalla Mura, *Senior Member, IEEE,* Jocelyn Chanussot, *Fellow, IEEE,* Christian Jutten, *Fellow, IEEE,* Paolo Gamba, *Fellow, IEEE,* and Andrea Marinoni, *Senior Member, IEEE*

***Abstract*—Although multimodal remote sensing data analysis can strongly improve the characterization of physical phenomena on Earth's surface, nonidealities and estimation imperfections between records and investigation models can limit its actual information extraction ability. In this paper, we aim at predicting the maximum information extraction that can be reached when analyzing a given dataset. By means of an asymptotic information theory-based approach, we investigate the reliability and accuracy that can be achieved under optimal conditions for multimodal analysis as a function of data statistics and parameters that characterize the multimodal scenario to be addressed. Our approach leads to the definition of two indices that can be easily computed before the actual processing takes place. Moreover, we report in this paper how they can be used for operational use in terms of image selection in order to maximize the robustness of the multimodal analysis, as well as to properly design data collection campaigns for understanding and quantifying physical phenomena. Experimental results show the consistency of our approach.**

***Index Terms*—Multimodal remote sensing, capacity, reliability, data analysis, image selection.**

I. Introduction

Recently, multimodal remote sensing has attracted the interest of several scientists, mainly because of the huge potential encompassed by the diversification of the data [1]–[9]. It is expected that multimodal remote sensing enhances the understanding of Earth's surface physical phenomena. Nonetheless, it has been proved, either in remote sensing [4], [5] or in other data science research fields [3], [10], that processing more data and records does not always result in detailed information extraction because of nonidealities, mismatches and estimation imperfections [5], [11], [12].

This effect is even more evident in a multimodal analysis set-up, as the differences in temporal, spatial, spectral, and radiometric resolutions might affect the correct alignment, labelling and reference of the records to be processed [1]–[3], [5], [8], [9]. For instance, the outcome of the 2009-2010 annual contest on remote sensing data fusion showed that the set of two optical images alone could ensure better change detection in a flooded area, than when combined with two SAR images [5]. This can be considered as an example pointing out that increasing the size and diversity of a dataset does not always imply an improvement in accuracy and robustness of the analysis.

Thus, investigations have been conducted to extract accurate and reliable information from a set of heterogeneous records [3], [4]. These studies are instrumental to understand the limits of multimodal remote sensing analysis, as well as to quantify the actual benefits that such an analysis can provide to the characterization of phenomena occurring on Earth's surface. In fact, the effectiveness of multimodal remote sensing information extraction varies depending on the assumptions on feature statistics and the knowledge of the sample distributions. In order to obtain the best information extraction performance from multiple images, it is expected that the multimodal estimation procedure would ideally weight the inputs coming from any considered dataset, so to make the best use of the relevant information collected by the different sensors. This would result in an adaptive scheme to achieve optimality in information extraction, but that might be too complex to implement without a complete knowledge of feature distributions and pixel noise statistics [13].

Understanding the maximum grade of information extraction (i.e., *capacity*) that can be achieved starting from a dataset with specific statistical properties plays a fundamental role in properly assessing the maximum performance of a multimodal analysis scheme. Furthermore, quantifying the reliability on the outcomes that a multimodal investigation architecture can produce is a key-factor in evaluating a value of the outputs [6], [11]–[14].

These quantities can have an operational use in several aspects of environmental science. In fact, they could enhance the selection of data subsets to achieve the maximum information extraction. In other terms, it would be possible to understand which images, features, or classifications are providing information on the considered scene, and retrieve the best subset from the remote sensing data pool. At the same

S. Chlaily and A. Marinoni are with Department of Physics and Technology, UiT the Arctic University of Norway, NO-9037 Tromsø, Norway (e-mail: {saloua.chlaily, andrea.marinoni}@uit.no).

M. Dalla Mura is with Grenoble Images Parole Signals Automatics Laboratory (GIPSA-lab), Department of Images and Signals, Grenoble Institute of Technology, F-38402 Grenoble, France, and with Tokyo Tech World Research Hub Initiative (WRHI), School of Computing, Tokyo Institute of Technology, Tokyo, Japan (e-mail: mauro.dalla-mura@gipsa-lab.grenoble-inp.fr).

J. Chanussot is with University of Grenoble Alpes, CNRS, Grenoble INP, GIPSA-lab, 38000 Grenoble, France (e-mail: jocelyn.chanussot@gipsa-lab.grenoble-inp.fr).

C. Jutten is with University of Grenoble Alpes, GIPSA-lab, 38000 Grenoble, France, and with Institut Universitaire de France, F-75005 Paris, France (e-mail: christian.jutten@gipsa-lab.grenoble-inp.fr).

P. Gamba is with Telecommunications and remote sensing laboratory, Department of Electrical, Computer, and Biomedical Engineering, University of Pavia, I-27100 Pavia, Italy (e-mail: paolo.gamba@unipv.it).

time, this result can lead to an enhancement of the system efficiency, since only the relevant features and attributes from the different images will be retrieved.

Furthermore, quantifying capacity and reliability of multimodal data would help in planning campaigns for pointwise relevant field investigation. Indeed, estimating a priori the reliability of our analysis on a given set of remotely sensed data over a specific area can help policy makers and end-users in verifying whether further missions (e.g., acquisition of more images, or man-driven exploration in specific region) would be required to characterize the phenomena of interest over the given region. Therefore, a strong improvement on the efficiency of data acquisition strategies (both in terms of financial costs and environmental impact) can be achieved by estimating the maximum information extraction that can be obtained from a new set of data. This last aspect is particularly important when extreme regions (e.g., polar areas, oceans, tropical forests) are analyzed since they typically require a strong effort in terms of manpower for data acquisition campaigns.

In this paper, we define the capacity, i.e., the maximum accuracy, that data analysis can achieve in a framework for multimodal remote sensing, as a function only of the images' statistics, and independently of the sample distributions. Our approach consists in considering the optimal adaptive weighting of the inputs able to achieve the maximum information extraction. This optimal set-up is obtained by an asymptotic investigation based on information theory. The asymptotic analysis enables the derivation of approximate expressions that can be measured before the multimodal processing takes place, and provides useful insight into the reliability of any outcomes. Thus, the main contributions of this paper are:

- the introduction of an information theory-based approach to characterize the maximum information extraction performance as a function of data statistics and parameters that characterize the multimodal scenario to be addressed;
- the definition of two indices to quantify maximum accuracy and reliability of the investigation strategy that can take place over the considered multimodal records;
- the use of an asymptotic investigation to approximate the expressions of the aforementioned indices, so that these quantities can be estimated during preprocessing, whilst the use of difficult numerical integration and root finding techniques can be avoided.

It is worth emphasizing that, to the best of our knowledge, this work represents the first effort devoted to the understanding of the information extraction process from multimodal remote sensing data, as this subject has not been addressed in this terms by any paper in technical literature. As such, in this paper we provide a thorough theoretical walk-through into the system set-up, the information theory-based definitions we have used and the approximations we considered. Moreover, in this manuscript we provide the evidence of the link between the capacity we can compute in preprocessing and the actual overall accuracy that can be obtained by completing the multimodal remote sensing data processing chain.

In the second part of this manuscript we aim at showing the actual potential of the aforesaid investigation and the proposed two indices for reliability and capacity of multimodal remote sensing data analysis. Specifically, we report the results achieved when setting two architectures (based on genetic algorithm and branch-and-bound scheme) aiming at the capacity maximization to automatically select at the data level the relevant information in the considered multimodal datasets. The achievements obtained when introducing the proposed image selection methods provide a solid overview of the actual impact that our information theory-based approach can have in unveiling the details of capacity and limits of information extraction in multimodal remote sensing applications, as well as in boosting the effective interpretation of the significant characteristics of the scenes.

The remainder of this paper is organized as follows. Section II reports the derivation of the capacity and the reliability metrics: Section II-A introduces the proposed model for multimodal data system and its motivation, and the details of the proposed information theory-based analysis are delivered in Sections II-B, II-C, and II-D with computation details in Appendix. In Section III, we introduce the methods for automatic image selection based on the asymptotic information indices that we have proposed. Then, Section IV delivers experimental performance results to show the consistency of our approach. Finally, Section V delivers our final remarks and some ideas on future research steps.

For notational convenience, random scalars are denoted by lower case letters, e.g., z. Underlined lower case letters designate random vectors, e.g., $\underline{z}$. The hat is used to denote estimates, e.g., $\hat{\underline{z}}$ refers to the estimate of $\underline{z}$. Double underlined upper case letters refer to matrices, e.g., $\underline{\underline{A}}$, $\mathbb{E}[\cdot]$ is the expectation operator, $\underline{z}^T$ identifies the transpose of $\underline{z}$ and $\underline{\underline{I}}_a$ the $a \times a$ identity matrix. Finally, $|\underline{\underline{A}}|$ is the determinant of matrix $\underline{\underline{A}}$.

II. Capacity and reliability of multimodal remote sensing

A. System model

For sake of clarity, we consider multimodal data fusion at the decision-level. In decision-level fusion, the estimates are drawn separately from each remotely sensed image or image set [4]. Then, the multimodal analysis combines these estimates to obtain more accurate and robust results. The results of this paper can be easily extended to other levels of fusion, i.e., feature-level and data-level [3], [4].

Let M be the number of the remotely sensed images to be processed, and L the numbers of pixels in the reference image. We denote by $\hat{\underline{x}}_{lm} = [\hat{x}_{lmr}]_{r=1,\dots,R}$ the set of decision outcomes of the l-th pixel drawn from the m-th image used to feed the multimodal analysis. The value and meaning of R and $\hat{\underline{x}}_{lm}$ would vary according to the final purpose of the considered analysis [1], [2], [6], [15]. For instance, in an unmixing framework, R would be the number of endmembers in the region, and $\hat{x}_{lmr}$ would be the abundance of the r-th endmember within the l-th pixel. In a supervised (unsupervised) classification scheme, R would be the number of classes (thematic clusters), and $\hat{x}_{lmr}$ would be the probability to be assigned to the r-th class of the l-th pixel in the region

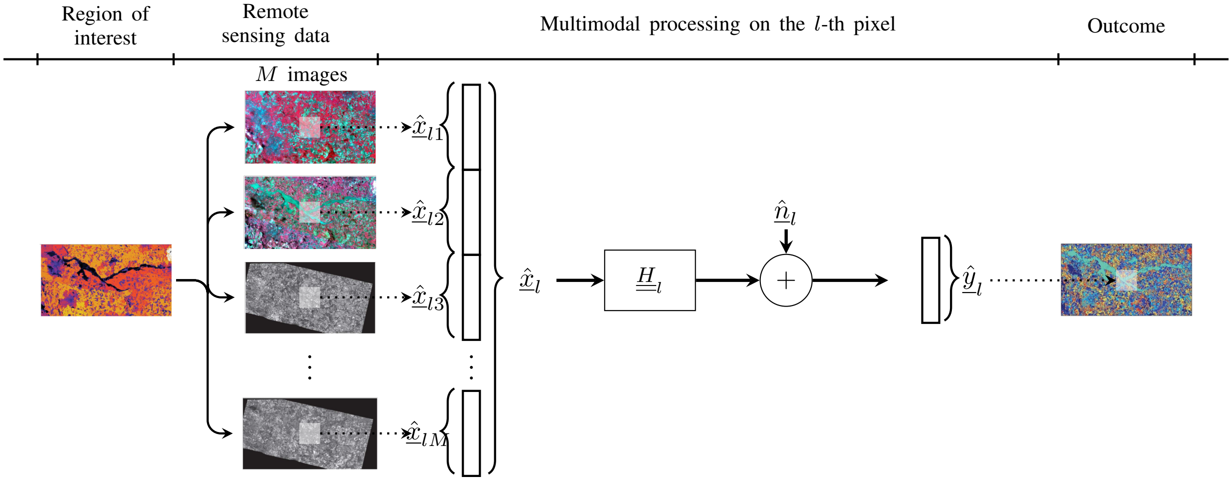

Region of interest
Remote sensing data
Multimodal processing on the l-th pixel
Outcome
M images
$\underline{\hat{x}}_{l1}$
$\underline{\hat{x}}_{l2}$
$\underline{\hat{x}}_{l3}$
$\underline{\hat{x}}_{lM}$
$\underline{\hat{x}}_{l}$
$\underline{\underline{H}}_l$
$\underline{\hat{n}}_l$
$+$
$\underline{\hat{y}}_l$

Fig. 1: General multimodal remote sensing analysis scheme that has been considered in this work. The multimodal processing is shown for the l-th pixel represented by grey squares in each remotely sensed image and also in the outcome of the analysis. Graphics are taken from the multimodal dataset delivered in the IEEE GRSS Data Fusion Contest 2009-2010 [5].

(thematic cluster). Finally, in target detection, R would be set to 1, and $\hat{x}_{lm}$ would then identify the likelihood of occurrence of the given target in the l-th pixel.

We write the estimates drawn on the l-th pixel of the region of interest out of the m-th image as $\underline{\hat{x}}_{lm} = f_{lm}(\underline{\underline{Y}})$, where $\underline{\underline{Y}} = [\underline{y}_l]_{l=1,\dots,L}$, with $\underline{y}_l = [y_{lr}]_{r=1,\dots,R}$, is the reference image. By reference image, we mean the information that we aim to recover on the observed scene. We can assume without losing generality that y_{lr} might live in a Q-level quantized $[0,1]$ interval, i.e., $y_{lr} \in \{i/(Q-1)\}_{i=0,1,\dots,Q-1}$. The f_{lm} functions identify different manipulations of the features acquired in the m-th image for properly addressing the mapping of different resolutions in terms of ground cover, geometry, radiometry, and spectral characterization (e.g., by filtering, convolving, integrating, or inferring).

Multimodal remote sensing combines the estimates $\underline{\hat{x}}_l$ in order to obtain a stable assessment of the actual $\underline{y}_l$ distribution for each pixel in the ground truth image. Specifically, $\underline{\hat{y}}_l \in \mathbb{R}^R$ is the outcome of the multimodal analysis, obtained by combining the estimates $\underline{\hat{x}}_l = [\underline{\hat{x}}_{lm}]_{m=1,\dots,M} \in \mathbb{R}^{RM}$,

$$\underline{\hat{y}}_l = \underline{\underline{H}}_l \underline{\hat{x}}_l + \underline{n}_l, \tag{1}$$

where $\underline{n}_l \in \mathbb{R}^R$ and $\underline{\underline{H}}_l \in \mathbb{R}^{R \times RM}$ model the imperfections and undesired mismatches generated through the estimation process. Although it may be more appropriate to model the effects of mismatches on multimodal remote sensing as nonlinear, a linear model enables the derivation of simple analytic expressions, which are more convenient to implement, analyze and interpret. Finally, we can assume that each element of $\underline{\hat{x}}_l$ lives in the Q-level quantized $[0,1]$ interval as the aforesaid elements of $\underline{\underline{Y}}$. Figure 1 reports a schematic workflow of the aforesaid model.

We consider every element $\underline{n}_l$ to be zero-mean Gaussian distributed. Moreover, to simplify the calculations, we state that $\mathbb{E}[\underline{n}_l^T \underline{n}_l] = N_0 \underline{\underline{I}}_R$ $\forall l$, where $N_0 \in \mathbb{R}_+$ is the noise variance. It is noteworthy that the model is still valid even if $\mathbb{E}[\underline{n}_l^T \underline{n}_l] \neq N_0 \underline{\underline{I}}_R$, since it suffices to multiply $\underline{\hat{y}}_l$ in (1) by the inverse of the noise covariance matrix yielding to the channel matrix $\left(\mathbb{E}[\underline{n}_l^T \underline{n}_l]\right)^{-1} \underline{\underline{H}}_l$. We will also assume $\underline{\underline{H}}_l$ to be stationary and ergodic, as well as to be composed of independent and identically distributed (i.i.d.) Gaussian random variables with zero mean. Moreover, we can assume $\underline{\underline{H}}_l$ to be slowly varying along the pixels of the reference map.

It is worth noting that these assumptions on the statistical properties of the terms in (1) are widely employed in the technical literature. Specifically, the assumptions of stationarity and ergodicity of the system either in terms of data statistics (e.g., [1], [16], [17], and [18] chap. 3) and processing approach (e.g., [2], [19]–[21], [18] chap. 16) are implied and/or drawn in order to maximize the trade off between precision and efficiency of the remote sensing data analysis frameworks. It is also true that in some specific application scenarios these assumptions (especially ergodicity) might not stand, particularly when specific applications with *a priori* known statistical properties in dynamic environments are targeted (e.g., [22]–[24]). Nonetheless, when no reliable *a priori* statistical knowledge can be specifically addressed and/or when facing the ambition to derive a general description for a multi-purpose data analysis scheme (as in the case of this work), the assumptions on (1) are able to provide a solid bedrock to achieve a good generalization of the system.

Furthermore, the statistical properties of $\underline{\underline{H}}_l$ and $\underline{n}_l$ might indeed follow several non-Gaussian distributions (e.g., Poisson-like, exponential), as nonidealities can show up along the multimodal remote sensing framework in several ways (e.g., detectors affected mainly by photon noise or images affected by striping effect). This paper aims at focusing on the definition of the optimal information extraction performance of a multimodal remote sensing data analysis framework in the case of most deteriorating imperfections. As such, it has been proven in [25] that Gaussian distributed imperfections would lead to a strong degradation of the information retrieval results compared to other imperfections.

It is possible to reshape the model in (1) through a singular value decomposition of $\underline{\underline{H}}_l$ [26], so that describing its information extraction properties can be facilitated. Let us define $\underline{\underline{W}}_l = \underline{\underline{H}}_l \underline{\underline{H}}_l^T \in \mathbb{R}^{R \times R}$. $\underline{\underline{W}}_l$ can have at most R non-zero eigenvalues as it is a Wishart matrix by definition [27]. Exploiting this property, we can decompose $\underline{\underline{H}}_l$ by means of a singular value decomposition as $\underline{\underline{H}}_l = \underline{\underline{U}}_l \underline{\underline{\Lambda}}_l \underline{\underline{V}}_l^T$, where $\underline{\underline{U}}_l \in \mathbb{R}^{R \times R}$ and $\underline{\underline{V}}_l \in \mathbb{R}^{RM \times RM}$ are unitary matrices, i.e., $\underline{\underline{U}}_l \underline{\underline{U}}_l^T = \underline{\underline{I}}_R$ and $\underline{\underline{V}}_l \underline{\underline{V}}_l^T = \underline{\underline{I}}_{RM}$. Moreover, $\underline{\underline{\Lambda}}_l \in \mathbb{R}^{R \times RM}$ has non-zero elements only on its main diagonal, *i.e.*, $\underline{\underline{\Lambda}}_{l_{rr}} = \sqrt{\lambda_{lr}}$ ($r \in \{1, \ldots, R\}$), and $\underline{\underline{\Lambda}}_{l_{rk}} = 0$ for $k \neq r$ ($k \in \{1, \ldots, MR\}$). $\sqrt{\lambda_{lr}}$ denotes the nonnegative square roots of the eigenvalues of $\underline{\underline{W}}_l$. Hence, by applying the aforesaid expression of $\underline{\underline{H}}_l$, we obtain the following alternative expression of (1),

$$\tilde{\underline{y}}_l = \underline{\underline{\Lambda}}_l \tilde{\underline{x}}_l + \tilde{\underline{n}}_l, \tag{2}$$

where $\tilde{\underline{y}}_l = \underline{\underline{U}}_l^T \hat{\underline{\underline{y}}}$, $\tilde{\underline{x}}_l = \underline{\underline{V}}_l^T \hat{\underline{x}}_l$, and $\tilde{\underline{n}}_l = \underline{\underline{U}}_l^T \underline{n}_l$. It is important to emphasize that, since $\underline{\underline{U}}_l$ and $\underline{\underline{V}}_l$ are unitary, the statistical properties of $\hat{\underline{x}}_l$ and $\underline{n}_l$, are maintained by $\tilde{\underline{x}}_l$ and $\tilde{\underline{n}}_l$, respectively. Namely, $\underline{n}_l$ and $\tilde{\underline{n}}_l$ have the same distribution and $\mathbb{E}[\hat{\underline{x}}_l^T \hat{\underline{x}}_l] = \mathbb{E}[\tilde{\underline{x}}_l^T \tilde{\underline{x}}_l]$.

Before introducing the method used to derive the capacity and reliability metrics, let us define and summarize in Table I some additional quantities that are crucial for our analysis.

TABLE I: Table of symbols.

$\hat{P}_{lr} = \frac{1}{M} \sum_{m=1}^{M} \hat{x}_{lmr}^2$	Average power of r-th component within the l-th pixel
$\hat{P}_l = \frac{1}{R} \sum_{r=1}^{R} \hat{P}_{lr}$	Average power of l-th pixel
$\hat{P} = \frac{1}{L} \sum_{l=1}^{L} \hat{P}_l$	Average power of all pixels
$\bar{\zeta} = \frac{\hat{P}}{RLN_0}$	Average SNR on the M images of all pixels
$\zeta_{lr} = \lambda_{lr} \frac{\hat{P}_{lr}}{N_0}$	SNR of the r-th component within the l-th pixel

B. Information theory-based analysis

The final purpose of an estimation scheme is to achieve the minimization of the variance between the target parameters and the produced estimates. This can be expressed, according to Cramer-Rao bound, in terms of the maximization of the mutual information between those two variables [11], [13], [28]. In estimation theory, this translates into quantifying how much of the target parameters can be reliably described by considering the set of observed variables. Hence, given θ the variable to be estimated and $\hat{\theta}$ its estimate, the mutual information can be expressed as $I(\theta, \hat{\theta}) = \sum_{\theta \in \Theta} \sum_{\hat{\theta} \in \hat{\Theta}} p(\theta, \hat{\theta}) \log \frac{p(\theta, \hat{\theta})}{p(\theta) p(\hat{\theta})}$, where $p(\theta, \hat{\theta})$ identifies the joint probability of θ and $\hat{\theta}$, whilst $p(z)$ reports the marginal probability of z. Further, the upper bound of $I(\theta, \hat{\theta})$ is called *capacity*, i.e., $\mathcal{C} = \sup_{p(\theta)} I(\theta, \hat{\theta})$. Therefore, the capacity determines the maximum value of information on θ that can be reliably extracted by the considered estimation system which provides $\hat{\theta}$ [13].

In this paper, we are interested in investigating the capacity of a multimodal remote sensing system. Thus, we must compute the upper bound of the mutual information between the estimate of the reference image $\hat{\underline{\underline{Y}}}$ based on the fusion of local estimates for all the M images and the set of estimates for each pixel $\hat{\underline{\underline{X}}} = [\hat{\underline{x}}_l]_{l=1,\ldots,L}$. Given the expressions in (1) and (2), it should not be surprising that the capacity for a multimodal remote sensing system depends on the images' statistical characteristics, as well as on the estimator performance, modeled by $\underline{n}$ and $\underline{\underline{H}}$. Nonetheless, when investigating the optimal maximum information extraction behaviour, the assumptions we have considered in the previous section are useful to find an analytical expression for the capacity of the multimodal system which is independent of the statistical distributions of the images.

Using the Gaussian assumptions (further details are given in appendix A), we define the capacity of the system in (1) as follows,

$$\mathcal{C} = \max_{\hat{P}_l} \sum_{l=1}^{L} \mathbb{E}_{\underline{\underline{H}}_l} \left[\log_Q \left| \underline{\underline{I}}_R + \frac{\underline{\underline{H}}_l \mathbb{E}[\hat{\underline{x}}_l \hat{\underline{x}}_l^T] \underline{\underline{H}}_l^T}{N_0} \right| \right], \tag{3}$$

where the maximum is over the average power of the l-th pixel, $\hat{P}_l = \mathbb{E}[\hat{\underline{x}}_l^T \hat{\underline{x}}_l]$, $\log_Q x = \frac{\ln x}{\ln Q}$ and Q refers to the quantization levels of $\hat{\underline{x}}_l$. The expression in (2) allows to simplify this equation to (see appendix A):

$$\mathcal{C} = \max_{\substack{\hat{P}_{lr} \\ \text{s.t.} \frac{1}{LR} \sum_{l,r} \hat{P}_{lr} = \hat{P}}} \sum_{l=1}^{L} \sum_{r=1}^{R} \mathbb{E}_{\zeta_{lr}} \left[\log_Q \left(1 + \frac{\zeta_{lr} \hat{P}_{lr}}{\hat{P}} \right) \right]. \tag{4}$$

where the maximum is taken over $\hat{P}_{lr}$ the average power (on the M images) of the r-th component within the l-th pixel, and ζ_{lr} is the SNR of the r-th component within the l-th pixel (see Table I).

The maximum information extraction can be achieved by an optimal power allocation that consists in attributing high power to more reliable estimates, and conversely, low, if not zero, power to unreliable estimates, whilst guaranteeing that the power constraint (i.e., $\frac{1}{LR} \sum_{l=1}^{L} \sum_{r=1}^{R} \hat{P}_{lr} = \hat{P}$) is still satisfied [13]. The following subsections provide more details on the terms of this maximization.

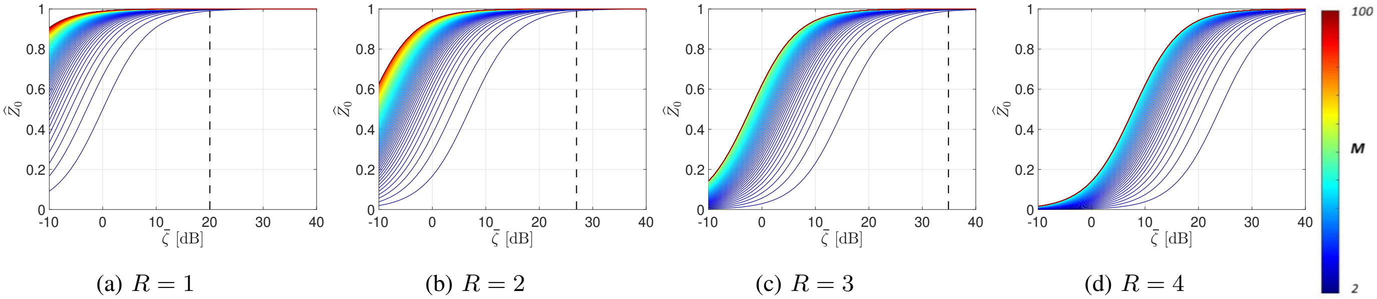
C. Reliability metric

The maximization in (4) is obtained by setting $\hat{P}_{lr}$, using a Lagrange multiplier, as follows [29]:

$$\frac{\hat{P}_{lr}}{\hat{P}} = \begin{cases} Z_0^{-1} - \zeta_{lr}^{-1} & \text{if } \zeta_{lr} \geq Z_0, \\ 0 & \text{if } \zeta_{lr} < Z_0, \end{cases} \tag{5}$$

where $r \in \{1, \ldots, R\}$, and Z_0 is the Lagrange multiplier.

According to the equation above, Z_0 can be considered as a cut-off value above which the retrieved estimates can be taken into account. In fact, if the estimates in $\hat{\underline{x}}_l$ show a signal-to-noise ratio (SNR) (defined according to the meaning of ζ_{lr}) smaller than Z_0, they should be discarded to ensure the maximum information extraction. Thus, Z_0 works as a metric that assesses the *reliability* of the estimates to provide the best understanding of the considered region. We can therefore ultimately use Z_0 to quantify the maximum confidence that we can achieve from the given multimodal framework [13].



(a) $R = 1$ (b) $R = 2$ (c) $R = 3$ (d) $R = 4$

Fig. 2: Reliability estimates obtained as in (11) for different settings of a multimodal remote sensing analysis framework. The color of the curves identifies the number of images M that are considered, according to the colormap on the right hand side of the figure. The number of classes/thematic clusters R varies from 1 to 4 in Figures (a) to (d), respectively. The trends of $\widehat{Z_0}$ are reported as a function of the signal-to-noise ratio (SNR) $\bar{\zeta}$. The dashed lines identify the minimum value of $\bar{\zeta}$ for which the maximum reliability is achieved by processing 2 images for different number of classes/thematic clusters R.

Since the value of $\hat{P}_{lr}$ varies with respect to ζ_{lr}, we write the discrete power constraint, i.e., $\frac{\sum_l \sum_r \hat{P}_{lr}}{LR} = \hat{P}$ as a continuous power constraint $\int \hat{P}_{lr} p_{\zeta_{lr}}(\zeta_{lr}) \mathrm{d}\zeta_{lr} = \hat{P}$. In order to derive the expression of Z_0, we substitute $\hat{P}_{lr}$ in the continuous power constraint by its definition in (5). Hence, Z_0 fulfills

$$\int_{Z_0}^{\infty} \left(Z_0^{-1} - \zeta_{lr}^{-1}\right) p_{\zeta_{lr}}(\zeta_{lr}) \mathrm{d}\zeta_{lr} = 1, \tag{6}$$

where $p_{\zeta_{lr}}(\zeta_{lr})$ is the probability distribution function of ζ_{lr}. Let us now consider the general distribution of the ζ_{lr} values over the whole dataset, i.e., $p_\zeta(\zeta)$. Thus, following the results in [14], (6) turns into

$$\int_{Z_0}^{\infty} \left(\frac{1}{Z_0} - \frac{1}{\zeta}\right) p_\zeta(\zeta) \mathrm{d}\zeta = 1. \tag{7}$$

Moreover, it is possible to define $p_\zeta(\zeta)$ by means of the properties of $\underline{\underline{W}}_l$. Indeed, recalling that $\underline{\underline{W}}_l$ is a Wishart distributed matrix [30], $p_\zeta(\zeta)$ can be written as, using that $\lambda = \zeta/\bar{\zeta}$,

$$p_\zeta(\zeta) = \bar{\zeta}^{-1} p_\lambda(\zeta/\bar{\zeta}), \tag{8}$$

where $\bar{\zeta} = \frac{\hat{P}}{RLN_0}$, is the average SNR, on all images, of all pixels (see Table I). $p_\lambda(\zeta/\bar{\zeta})$ represents the probability distribution function of an unordered eigenvalue of a Wishart distributed matrix [13], [14], [26], [31], i.e.,

$$p_\lambda(u) = \frac{e^{-u} u^{R(M-1)}}{R} \sum_{r=1}^{R} \kappa_r \left[\mathcal{L}_{r-1}^{R(M-1)}(u)\right]^2, \tag{9}$$

being $\mathcal{L}_s^t(u) = \frac{1}{s!} e^u u^{-t} \frac{\mathrm{d}^s}{\mathrm{d}u^s}\left(e^{-u} u^{s+t}\right)$ the s-th order Laguerre polynomial and $\kappa_r = \frac{(r-1)!}{(RM-R+r-1)!}$.

It is possible to prove that Z_0 is living in $[0,1]$, $\lim_{\bar{\zeta}\to 0} Z_0 = 0$, and $\lim_{\bar{\zeta}\to\infty} Z_0 = 1$ [14]. Moreover, (7) (and thus (4)) has a unique solution in Z_0. Hence, there is only one value of Z_0 able to provide the maximization of the information extraction in the multimodal system in (1). In appendix B, we detail the proof of the uniqueness of Z_0.

The expression of Z_0 can be obtained by working on the algebraic properties of the terms involved in (7), as provided in Appendix C. In fact, (7) can be written as, with $\nu = Z_0/\bar{\zeta}$,

$$\sum_{r=1}^{R} \kappa_r \sum_{j_1=0}^{r-1} \sum_{j_2=0}^{r-1} \frac{(-1)^{j_1+j_2}}{j_1! j_2!} J_1 J_2 \frac{\Gamma(\beta_1,\nu) - \nu\Gamma(\beta_2,\nu)}{\nu} = R\bar{\zeta}, \tag{10}$$

where $J_1 = \binom{R(M-1)+r-1}{r-1-j_1}$ and $J_2 = \binom{R(M-1)+r-1}{r-1-j_2}$, $\beta_2 = R(M-1) + j_1 + j_2$, $\beta_1 = \beta_2 + 1$, and $\Gamma(s,t)$ is the complementary incomplete Gamma function, i.e., $\Gamma(s,t) = \int_t^\infty u^{s-1} e^{-u} \mathrm{d}u$. The above equation can be solved for a unique value of the cut-off Z_0 by means of iterative numerical search techniques. Nevertheless, it is possible to obtain an approximate estimate of Z_0 by means of a small perturbation procedure applied to ν in (10). Let us consider $\alpha(t) = \sum_{r=1}^{R} \kappa_r \sum_{j_1=0}^{r-1} \sum_{j_2=0}^{r-1} \frac{(-1)^{j_1+j_2}}{j_1! j_2!} J_1 J_2 t!$. Hence, it is possible to write the approximate of Z_0 as

$$Z_0 \approx \widehat{Z_0} = \frac{\bar{\zeta}\alpha(\beta_2)}{R\bar{\zeta} + \alpha(\beta_2 - 1)}, \tag{11}$$

It is worth noting that, as a first result of the aforesaid definition, (11) becomes an algebraic equation for $\bar{\zeta} \gg 1$, i.e., $\lim_{\bar{\zeta}\to\infty} \widehat{Z_0} = \frac{1}{R}\alpha(\beta_2) = 1$.

Now, to understand how the reliability metric reacts to various set-ups of multimodal remote sensing data analysis framework, using equation (11), we report in Fig. 2 the variation of $\widehat{Z_0}$ as a function of $\bar{\zeta}$ and for different values of M and R. We remark that $\widehat{Z_0}$ trends change significantly for different set-ups. For a given M and R, $\widehat{Z_0}$ increases with the average SNR, which is expected since the extracted information from the images gets more precise. Moreover, while comparing the subfigures, we notice that the curves are shifted to the right when the number of classes/thematic clusters R increases, showing that the reliability decreases when R increases. For example, the optimal information extraction is reached with maximum reliability (when $M = 2$) for $\bar{\zeta}$ equal to 20 dB, 27 dB, 35dB, and more than 40 dB for $R = 1, \ldots, 4$, respectively. In fact, by increasing the number of classes/thematic clusters, the complexity of the analysis increases. In this case, more data and better SNR are required to extract further details about the observed region. Finally, for a given $\bar{\zeta}$ and R, the reliability increases with the number of images, since additional information can be extracted. However, increasing the number of images is not always relevant. For instance, the curves for $M > 35$ collapse when R equals 3 and 4. In these cases, the optimal information extraction is achieved with fewer images, and additional data does not provide any novelty.

D. Capacity metric

Let us focus now on the maximization of (4). Given the optimization settings in (5) and the aforementioned results on the cut-off value Z_0, the capacity of a multimodal system as in (1) can be written as follows,

$$\mathcal{C} = LR \int_{Z_0}^{\infty} \log_Q \left(\frac{\zeta}{Z_0} \right) p_\zeta(\zeta) \mathrm{d}\zeta, \tag{12}$$

where $p_\zeta(\zeta)$ is the probability distribution function of any eigenvalue in the system as in (8). It is possible to write an approximation of (12) in a consistent way with respect to (10), following [14], [32],

$$\mathcal{C} \approx \widehat{\mathcal{C}} = \sum_{r=1}^{R} \kappa_r \sum_{j_1=0}^{r-1} \sum_{j_2=0}^{r-1} \frac{(-1)^{j_1+j_2}}{j_1! j_2!} J_1 J_2 \mathcal{B}_{\beta_1}(\nu). \tag{13}$$

where,

$$\mathcal{B}_t(\nu) = (t-1)! \left[E_1(\nu) + \sum_{k=1}^{t-1} \mathcal{P}_k(\nu)/k \right], \tag{14}$$

where $E_1(\nu) = \int_\nu^\infty e^{-u} u^{-1} \mathrm{d}u$ is the exponential integral function, whereas $\mathcal{P}_k(\nu) = e^{-\nu} \sum_{m=0}^{k-1} \nu^m / m!$ is the k-th order Poisson sum. Moreover, let us recall that $\beta_1 = R(M-1) + j_1 + j_2 + 1$.

$\mathcal{C}$ reports the number of quantization levels that can be reliably analyzed and processed over the given set of multimodal images per areal unit. Hence, it provides a quantification of the *precision* that can be reached when analyzing the considered region, i.e., the degree of understanding that can be extracted on the area of interest. This approximation $\widehat{\mathcal{C}}$ in (13) describes the level of detail reachable if the given multimodal analysis framework works optimally. $\widehat{\mathcal{C}}$ is a function of the local signal-to-noise ratio computed over the images and estimates. Hence, by computing the SNR of the given remote sensing data pool, the information extraction limits for any multimodal dataset can be calculated.

In Fig. 3, we show capacity curves as function of the average SNR $\bar{\zeta}$ according to equation (13). The different subfigures correspond to different number of classes/thematic clusters R and the curves' colors refer to the number of used images M. Note that the scale of the vertical axis is different for each subfigure. It is not surprising that the precision of the optimal information extraction[†] increases with the number of classes/thematic clusters R, with the average SNR, and with the number of images M. In fact, in these cases, further details about the region of interest can be obtained. However, the gain in capacity decreases by increasing the number of images until it is saturated, and no improvement can be obtained, as can be seen in the zoomed part of subfigure 3d. In fact, compared to two images, 12 images provide a capacity gain of Δ, which represents half of the gain obtained by adding 98 images. Moreover, we notice that curves with low M deliver a higher value of capacity, especially for higher R and lower $\bar{\zeta}$. Low values of M correspond to the case where a small amount of images is considered. This result might seem counterintuitive since it is commonly presumed that in the case of a noisy dataset, more data ensures better results. However, for the data fusion to be fruitful, the multimodal analysis should be able to capture the underlying structure of the images. This task becomes difficult when increasing the size of the dataset and even more complicated when the considered dataset is noisy. The capacity metric is able to determine the threshold on the average SNR over which more data actually implies better results. Note that the value of such threshold increases with the complexity of the analysis (i.e., with R).

III. Automatic image selection

As mentioned, the metrics derived in Section II can be used to perform a selection of the relevant contributions within the considered datasets at data, feature, and decision level. In this paper, we focus on building a framework for automatic image selection, i.e., on designing an architecture that selects the most informative scenes in the considered dataset in order to perform the best information extraction. It is worth emphasizing that this can be considered as a straightforward application of the notions that we addressed in Section II. Indeed, other applications could be also considered. Specifically, future works will then be dedicated on the use of $\widehat{Z}$ and $\widehat{\mathcal{C}}$ for enhanced feature selection and ensemble learning, when combining the decisions drawn from a set of different classifiers on the same set of records.

Selecting the most informative subset of images identifies a complicated optimization problem, that can be generally described as follows:

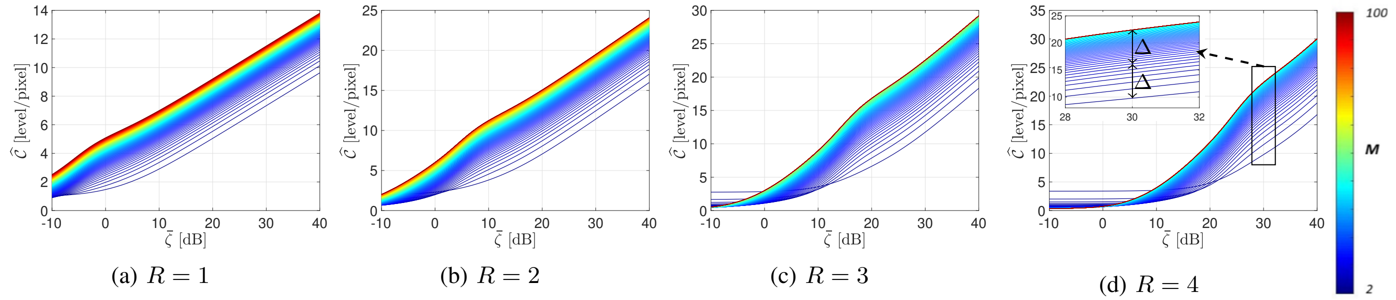
$$\arg\max_{x \in \mathcal{D}} \widehat{\mathcal{C}}(x), \tag{15}$$

where $\mathcal{D}$ identifies the set of all possible combinations of the images in the dataset and $\widehat{\mathcal{C}}(x)$ is the value obtained when all the quantities in (13) are computed over the subset x.

Indeed, the goal that we would like to achieve directly translates into optimizing the number of images to be considered, as well as the actual composition of the selected subset. As such, it is possible to expect that the optimization problem we must face is nonlinear and eventually non-convex [30], [33]. Therefore, in general, this problem might have multiple local solutions. The number of such local solutions is not known a priori, and the quality of global and local solutions may differ substantially [33].

Thus, it is not possible to directly use standard optimization strategies to solve this problem, and decision models can be very difficult to apply [30], [33]. In fact, classic local search methods might be very sensitive to the initial state of the optimization procedure, which can strongly affect the quality of the search process. Hence, a global search method

[†] It is measured in the number of levels that can be reliably solved for an areal unit



(a) $R = 1$

(b) $R = 2$

(c) $R = 3$

(d) $R = 4$

Fig. 3: Capacity estimates according to (13) for different settings of a multimodal remote sensing analysis framework. The same legend as in Fig. 2 applies here. For better visibility, the scale of the vertical axis is different for each subfigure.

must be considered. Several global search methods have been introduced in data analysis and remote sensing processing (see for instance [6], [12], [30], [33]–[36]). In particular, global search methods can be categorized as belonging to two main classes, i.e., exact methods and heuristic methods.

Exact methods aim at directly solving the problem by exhaustively enumerating all possible solutions [37], [38]; visiting all stationary points of the objective function leading to the list of all optima [39]; replacing the problem through a successive refinement process, by a sequence of relaxed subproblems that are easier to solve [38]; allowing a stochastic description of the modeled function/class, so that the characteristics of problem/instance are adaptively estimated and updated during optimization [40]; random sampling within the feasible sets to perform adaptive stochastic search [37]; approximating the level sets of the objective function to determine its supremum [41].

A different approach to achieve global optimization is represented by heuristic algorithms. This family of methods tackle the nonlinearities involved in the feasible set by randomly choosing multiple initial states, so to trigger several search processes and choose the best solution [37]; starting from an initial solution, performing a set of steps that lead to a new solution inside the neighborhood of the current one while forbidding to move to points already visited in the search space (e.g., tabu methods) [42]; investigating the convexity characteristics of the feasible sets, and then performing direct sampling to reach global optimization [43]; transforming the objective function into a smoother function with easily calculated local minimizers, and then using a local minimization procedure to trace all minimizers back to the original function [44]; building auxiliary functions to assist the searching procedure (e.g., tunneling, deflation, and filled function approaches) [45].

In order to perform a solid comparison of the benefits and drawbacks of these algorithms when applied to the maximization of the information extraction, we chose to consider one representative for each of these two families of optimization methods. Specifically, we implemented image selection using a branch-and-bound approach and genetic optimization.

Branch-and-bound (BB) algorithms can be considered as strategies based upon adaptive partition, sampling, and bounding procedures to solve global optimization. These operations are applied iteratively to the collection of active subsets within the feasible set. BB algorithms are therefore exact methods which typically rely on some a priori structural knowledge about the problem such as how rapidly each function can vary (e.g., the knowledge of a suitable overall Lipschitz constant). This general methodology is applicable to a large number of global optimization problems (such as concave minimization or reverse convex programs [37], [41]). It can be considered as a valid candidate for the practical implementation of the automatic image selection we would like to achieve.

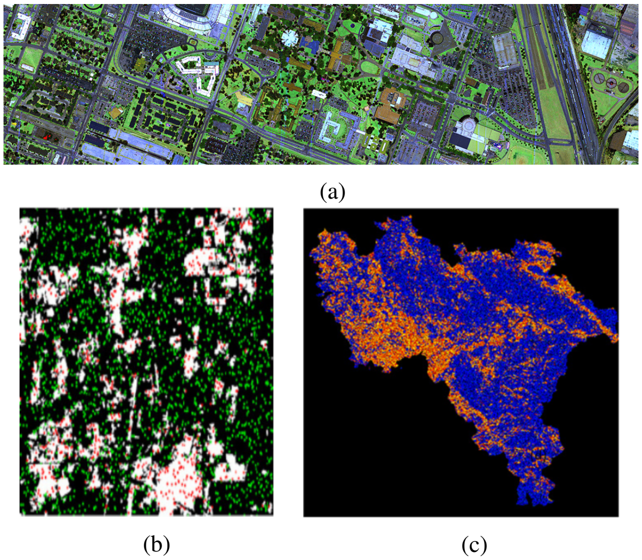
When applying BB to solve the problem in (15), the overlapping version of BB algorithm (OBB) [46] can be properly applied. To this aim, the bounds of the search procedure are computed by means of nonconvex estimators of the $\widehat{\mathcal{C}}(x)$ function for the considered combination of images x. In practice, the feasible set $\mathcal{D}$ of all the image combinations is iteratively covered by focusing and refining balls $\mathcal{B}$ (i.e., subspaces of $\mathcal{D}$ that are covering multidimensional regions equidistant from their centroid) that are partially overlapping in the space induced by $\mathcal{D}$. The radius of $\mathcal{B}$ is adaptively but monotonically decreased with the iteration progress. This step achieves branching, while the bounding phase is performed by computing lower and upper bounds of $\widehat{\mathcal{C}}(x)$ over the subregion $\mathcal{B}$ as $\max\{0, \widehat{\mathcal{C}}(x_{\mathcal{B}}) \mp L_{\widehat{\mathcal{C}}}(\mathcal{B})\delta_{\mathcal{B}}\}$, where $x_{\mathcal{B}}$ is the closest point of $\mathcal{D}$ to the centroid of $\mathcal{B}$, $\delta_{\mathcal{B}}$ is its distance to the optimization point x, and $L_{\widehat{\mathcal{C}}}(\mathcal{B})$ is the Lipschitz constant of $\widehat{\mathcal{C}}(x)$ over $\mathcal{B}$.

On the other hand, genetic algorithms (GAs) heuristically simulate biological evolution [37], [40]. Based on a population of candidate solution points (in (15), $\mathcal{D}$), an adaptive search procedure is applied. The poorer solutions are dropped by a competitive selection, while the remaining pool of candidates with higher values are then recombined with other solutions by swapping components. Based on diverse evolutionary criteria, a variety of deterministic and stochastic algorithms can be constructed. Hence, these architectures can be used to solve the proposed problem by properly defining its structural requirements [40], [42].

It is worth pointing out that a detailed explanation of the theoretical aspects of the optimization schemes employed in this work is beyond the scope of this paper. For additional information on the OOB and GA approaches the reader is encouraged to refer to [46] and [40], respectively.

IV. Experimental results

The proposed approach has been tested on multiple datasets of remotely sensed records, acquired by optical and synthetic


(a)

(b)

(c)

Fig. 4: A few samples of the test datasets: IEEE GRSS Data Fusion Contest 2018 data (a), Beijing dataset (b), and Pavia air quality, glycated hemoglobin and BMI dataset (c).

aperture radar (SAR) sensors.

A. *Test datasets*

In this paper, we have tested our approach on three different multimodal scenarios, from datasets remotely acquired by multiple selected sensors, to multitemporal stacks of remote sensing records, to a collection of data from multiple sources, spaceborne and *in situ* (see Fig. 4).

First, we considered the dataset used for the IEEE GRSS Data Fusion Contest 2018, denoted DFC2018 in the following [47]. It consists of a collection of remotely sensed data acquired over the University of Houston, TX, on Feb. 16, 2017, covering the University of Houston campus and its surrounding areas. The remote sensing records are a multispectral-LiDAR point cloud data at 1550 nm, 1064 nm, and 532 nm; a hyperspectral dataset covering the 380-1050 nm spectral range with 48 bands at a 1-m ground sample distance (GSD); and a very high resolution RGB imagery at a 5-cm GSD. 20 classes have been identified. Thus, the values of (R, M) according to the notation in the previous section have been set to $(20, 3)$. It is noteworthy that DFC18 used in our analysis consists only of the training dataset in [47]. As such, our results cannot be compared to other publications that use the test dataset.

Then, we considered a dataset of 53 SAR images acquired over Beijing, People's Republic of China, by the ESA Sentinel-1 sensor from February 2015 to March 2016 [48]. Each scene consists of 3265×2448 VV and VH intensity records, where V and H identify the orientation of the polarization (vertical and horizontal, respectively). These records are used to identify regions that could be associated with urban and non-urban classes through time. Therefore, the values of R and M were set to 2 and 53, respectively.

Finally, we took into account a dataset that results from gathering together multispectral records collected by Landsat over the Province of Pavia, Italy from 2009 to 2014, and yearly and municipality-aggregated glycated hemoglobin and body mass index (BMI) values collected out of a group of 1084 patients affected by diabetes living in the area during the same time interval [15], [49]. Specifically, we collected 35 Landsat images, each consisting of 2800 × 2800 pixels at 30m spatial resolution. Only the inverse of the thermal infrared band of each Landsat image was used to retrieve the quantity of black particulate concentrated over the area [49]. Specifically, 5 air quality classes (designed according to the limits and specifics provided by the local environmental protection agency to determine the risk of diabetes-related complication onsets, validated by means of black particulate matter sensing ground stations and clinical records [49]) have been considered, i.e., $R = 5$. Moreover, the number of images (M) was set to 35.

TABLE II: Overall accuracy (OA) and capacity $\widehat{C}$ obtained when considering different subsets of images (M') and values of $\bar{\zeta}$ in dB on the Beijing dataset. OA is computed according to the classification performed by means of the method proposed in [48], whilst the capacity is computed as in (13). μ and std represent the average and standard deviation, respectively, of the OA obtained over 100 experiments with multiple random subsets of images.

	$M' = 5$			$M' = 15$			$M' = 25$		
	OA			OA			OA		
$\bar{\zeta}$	μ	std	$\widehat{C}$	μ	std	$\widehat{C}$	μ	std	$\widehat{C}$
2	**0.61**	0.04	**2.5**	0.53	0.045	2.1	0.52	0.06	2.1
10	0.71	0.052	5.8	0.79	0.031	7.8	**0.86**	0.036	**8.2**
20	0.74	0.057	9.6	0.85	0.021	11.3	**0.88**	0.013	**12.1**

TABLE III: Overall accuracy (OA) and capacity $\widehat{C}$ obtained when considering different subsets of images (M') and values of $\bar{\zeta}$ in dB on the Pavia dataset. The same legend as in Table II applies here.

	$M' = 3$			$M' = 13$			$M' = 23$		
	OA			OA			OA		
$\bar{\zeta}$	μ	std	$\widehat{C}$	μ	std	$\widehat{C}$	μ	std	$\widehat{C}$
5	**0.62**	0.049	**4.8**	0.52	0.048	2.3	0.51	0.045	2.2
10	**0.69**	0.015	**5**	0.67	0.014	4.95	0.67	0.0143	4.95
20	0.73	0.016	5.3	0.79	0.013	10.1	**0.81**	0.01	**11.3**

B. *Validation of $\widehat{Z_0}$ and $\widehat{C}$*

Fig. 5 shows the results in terms of reliability (a) and capacity (b) for these datasets. We computed the trends of $\widehat{Z_0}$ and $\widehat{C}$ as a function of $\bar{\zeta}$ according to (11) and (13), respectively, and for different values of R and M in the dataset, i.e., the number of classes and images considered in each dataset. These results are displayed in solid lines. Then, we computed the exact Z_0 by searching for the value that maximizes (7), and the exact value of $\mathcal{C}$ according to (12). These results are shown as stars. Moreover, in order to investigate the actual effectiveness of our approach, we added white Gaussian noise to the records, and then computed Z_0 and $\mathcal{C}$ as in (7) and (12) for different values of $\bar{\zeta}$ (dashed lines).

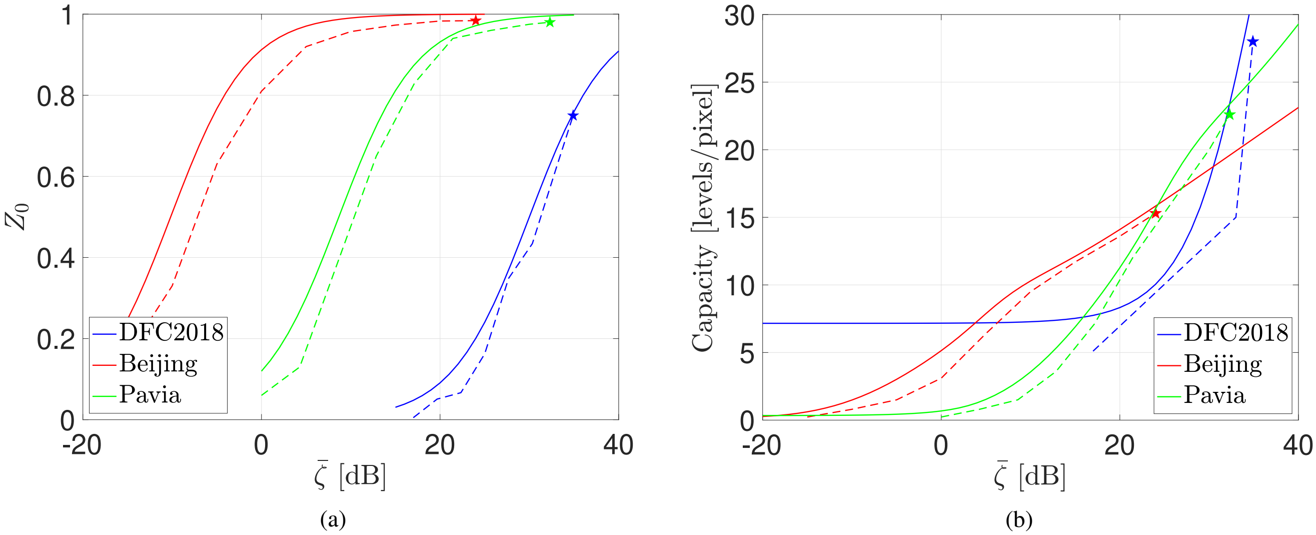


Fig. 5: Reliability and capacity estimates (in (a) and (b), respectively) obtained when considering the IEEE GRSS Data Fusion Contest 2018 dataset (blue lines), the Beijing SAR records (red), and the Pavia data (green). The solid lines identify the bounds computed as in (11) and (13). The star markers identify the exact values in (7) and (12) obtained on the actual datasets, whereas the dashed lines show these results obtained on the same datasets where we added white Gaussian noise. These trends are all displayed as a function of $\bar{\zeta}$.

These tests would help us to understand and quantify the actual ability of our approach to approximate the maximum values of reliability and capacity that a multimodal system can reach.

For all the datasets the approximations (11) and (13) are able to top the actual values of Z_0 and $\mathcal{C}$ that have been computed. This result is compliant with the assumptions and the goals of the proposed approach. Hence, we can state that the introduced metrics are able to describe the maximum values of reliability and accuracy of the information extraction process for a given multimodal dataset under the conditions described in Sections II-A and II-B. Moreover, these figures show that the derived approximate cutoff values are reasonable when $\bar{\zeta}$ is sufficiently large with respect to the actual values of Z_0 computed by the exact equation. Furthermore, they display that, although the cutoff estimate in (11) deviates from the true Z_0, it still results in a reasonable capacity estimate approximation. It is also worth noting that the approximate and exact trends are typically following the same behaviour for both Z_0 and $\mathcal{C}$ on any considered datasets. We would like to emphasize that the reasons why the DFC2018 dataset achieves the highest capacity even though it has the lowest reliability is merely related to the different multimodal set-up. As reported in Fig. 2 and Fig. 3, the reliability decreases with the number of classes/clusters (equal to 20 for DFC2018, and equal to 2 for Beijing and 5 for Pavia datasets), contrary to the capacity.

C. $\widehat{\mathcal{C}}$ and overall accuracy

Let us now compare the capacity index and overall accuracy (OA) of classification outcomes on Beijing and Pavia datasets. For each dataset, let us randomly identify a subset of M' images. Specifically, M' is set to $\{5, 15, 25\}$ for the Beijing dataset, whilst $M' = \{3, 13, 23\}$ when the Pavia dataset is considered. To these subsets, we added noise as in the previous experiment so to obtain different values of $\bar{\zeta}$. In fact, the value of $\bar{\zeta}$ is equal to $\{2, 10, 20\}$ (in dB) when we considered the Beijing datasets, whilst $\bar{\zeta} = \{5, 10, 20\}$ (in dB) for the Pavia dataset. Hence, we take into account nine subsets for both datasets characterized by a different set-up of the $(M', \bar{\zeta})$ parameters. Let us now perform classification on these sets of records. When considering the Beijing dataset, we are interested in identifying the urbanized area in the region. On the other hand, when we focus our attention on the Pavia records, we aim at understanding the level of air quality of each pixel. Then, let us compute the OA of the resulting classification according to ground truth maps acquired by ground measurements, as well as the estimated value of capacity according to (13). The OA is obtained over 100 experiments, for each set-up $(M', \bar{\zeta})$, with different random subsets of images. In order to provide a fair comparison, we took into account only one classifier, specifically the one introduced in [48]. It is worth recalling that the OA achieved by this classifier on the original dataset (i.e., without additive noise and using all images, so that $(\bar{\zeta}, R, M) = (24, 3, 53)$ for the Beijing dataset, and $(\bar{\zeta}, R, M) = (32, 5, 35)$ for the Pavia dataset) are equal to 94 $\%$ and 90.6 $\%$, respectively. Table II and III report the results on the aforementioned subsets.

When considering OA, it is typically expected that it would increase for a given value of SNR when the number of images (M') to be processed would increase as well. The results show that this effect is confirmed for higher values of SNR (e.g., $\bar{\zeta} = 20$dB for Beijing dataset), whilst the trend becomes the opposite for low-SNR set-up (i.e., $\bar{\zeta} = 2$dB for Beijing dataset). According to Tables II and III, the capacity index proposed in (13) is able to give an indication of this behavior.

The capacity index is especially useful to detect the inversion of the trend in OA in the low-SNR situation, where processing a fewer number of images delivers better results than performing classification with more scenes. According to these results, it is possible to confidently state that the assumptions taken into account in the system set-up and throughout the proposed information theory-based analysis are accurate and match the actual data processing structure, so that an operational use of the proposed indices can be actually developed and implemented.

D. Improved classification by enhanced automatic image selection

To validate the effectiveness of using the capacity metric for automatic selection of the most informative images in a multimodal remote sensing dataset, we performed several tests on the records presented in Section IV-A. Specifically, we implemented the OBB- and GA-based image selection as introduced in Section III: throughout this section these methods will be called $\widehat{\mathcal{C}}$-OBB and $\widehat{\mathcal{C}}$-GA, respectively.

In order to evaluate the actual performance improvement induced by the use of the $\widehat{\mathcal{C}}$ metric and understanding the relevance of the images, we performed automatic selection by means of other algorithms widely used in remote sensing technical literature based on different strategies and philosophy. Hence, we considered the following architectures:

- OBB and GA algorithms when the objective function was set according to the Mahalanobis criterion [37], [42], [46];
- forward feature selection (FS) based on the 1-nearest neighbor leave-one-out criterion [50];
- principal component analysis (PCA - see for instance [51]), based on eigenvalue analysis and ranking;
- discriminant boundary feature extraction (DBFE) [52], based on separability enhancement;
- Fisher information maximization (Fisher information scoring - FIS [53]).

These schemes are generally used at data level for multimodal remote sensing processing.

All these algorithms are then compared in terms of classification accuracy. In order to provide a fair comparison of the different schemes, we implemented a common platform for classification based on two algorithms, support vector machine (SVM) and random forest (RF), widely used by the remote sensing community (see for instance [54] and [55]). This step is meant on the one hand to provide a shared platform for the aforementioned schemes of image selection, so that consistent evaluation of the performance can be conducted; on the other hand, it gives also the opportunity to quantitatively address the level of applicability of the image selection architectures to different frameworks of remote sensing image classification. When RF is employed, the number of decision trees was empirically set to 100 decision trees, and $\lfloor\sqrt{F}\rfloor^{\ddagger}$ variables are randomly drawn at each node of the trees, where F is the number of images associated with each pixel that have been identified by the aforementioned selection architectures [56], [57]. SVMs exploit radial basis functions [58], with parameters C (penalty parameter of error term) and γ (kernel coefficient) automatically selected via cross-validated grid-search optimization [18].

TABLE IV: Definition of the classification performance indices

Index	Acronym	Definition
Accuracy	ACC	$\frac{TP+TN}{TP+FN+TN+FP}$
Sensitivity	TPR	$\frac{TP}{TP+FN}$
Specificity	TNR	$\frac{TN}{FP+TN}$
Precision	PPV	$\frac{TP}{TP+FP}$
F1 score	F1	$\frac{2TP}{2TP+FP+FN}$

TABLE V: Quantitative analysis of the tests conducted on Beijing dataset. For each index, the best result is shown in red, whilst the second is shown in blue. M_* represents the optimal number of selected attributes by each method.

Method	M_*	Classifier	ACC	TPR	TNR	PPV	F1
PCA	33	RF	0.83	0.87	0.84	0.85	0.86
		SVM	0.83	0.83	0.83	0.88	0.85
DBFE	34	RF	0.82	0.85	0.78	0.84	0.84
		SVM	0.84	0.84	0.84	0.77	0.85
FIS	**26**	RF	0.82	0.87	0.75	0.84	0.85
		SVM	0.82	0.82	0.82	0.88	0.85
FS	30	RF	0.8	0.77	0.79	0.78	0.81
		SVM	0.81	0.81	0.77	0.79	0.81
OBB	35	RF	0.76	0.8	0.79	0.78	0.79
		SVM	0.8	0.78	0.76	0.8	0.8
GA	34	RF	0.79	0.76	0.75	0.79	0.81
		SVM	0.74	0.74	0.8	0.76	0.74
$\widehat{\mathcal{C}}$-OBB	**27**	RF	**0.95**	**0.95**	**0.95**	**0.98**	**0.97**
		SVM	0.92	0.92	0.92	**0.98**	**0.95**
$\widehat{\mathcal{C}}$-GA	**26**	RF	0.91	0.93	0.87	0.95	0.94
		SVM	**0.94**	**0.94**	**0.94**	**0.97**	**0.95**
No slct.	53	RF	0.92	**0.95**	0.88	0.94	0.94

In summary, for each dataset we have produced 8 feature subsets identified by means of methods based on PCA, DBFE, FIS, FS, OBB and GA at data level, as well as by the two proposed schemes $\widehat{\mathcal{C}}$-OBB and $\widehat{\mathcal{C}}$-GA introduced in Section III. Moreover, we have used two different classifiers to evaluate the accuracy performance with respect to classification of these methods.

1) Beijing dataset: Let us focus first on the results obtained by processing the Beijing dataset. Since the classification problem on this dataset is essentially binary, we can evaluate the performances of the different methods according to the performance indices listed in Table IV, where the counts of true positives, true negatives, false positives and false negatives are listed as TP, TN, FP and FN, respectively. Table V reports the results achieved. It is possible to appreciate how the image selection based on the capacity metric (either conducted according to OBB or GA) provides a strong enhancement in all the accuracy indices with the least number of selected images.

To give a complete quantitative assessment of the actual impact of the results, we have reported the results achieved when

‡ $\lfloor x\rfloor$ denotes the floor function that returns the integer part of a real number x such that $\lfloor x\rfloor \leq x$

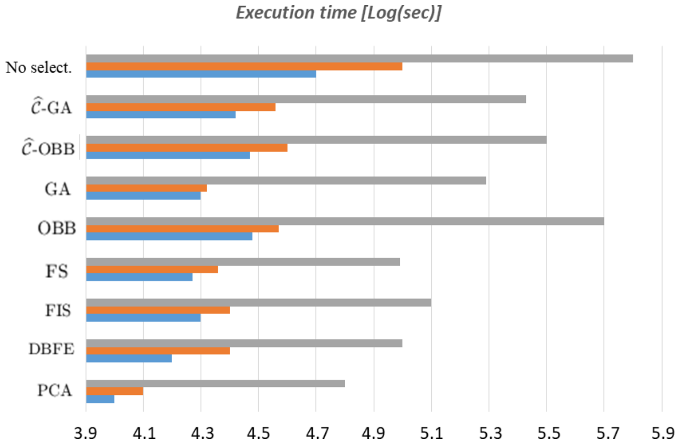


Fig. 6: Execution time (in $\log(\text{seconds})$) required by the architectures in Tables V-XIV (image selection and RF) to achieve the classification of Beijing, Pavia, and IEEE GRSS Data Fusion Contest 2018 datasets (blue, orange, and gray bars, respectively).

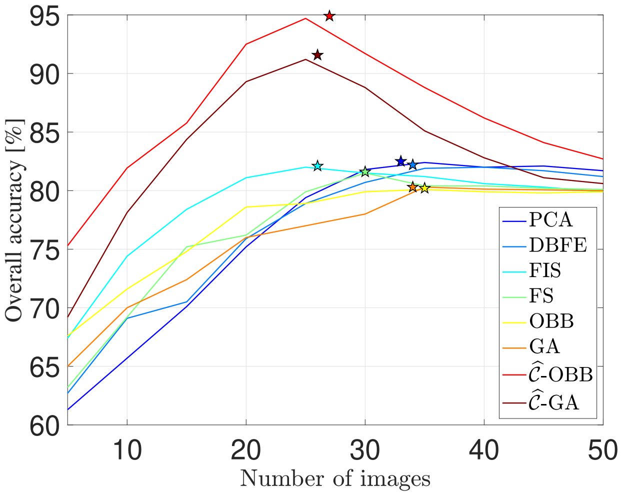


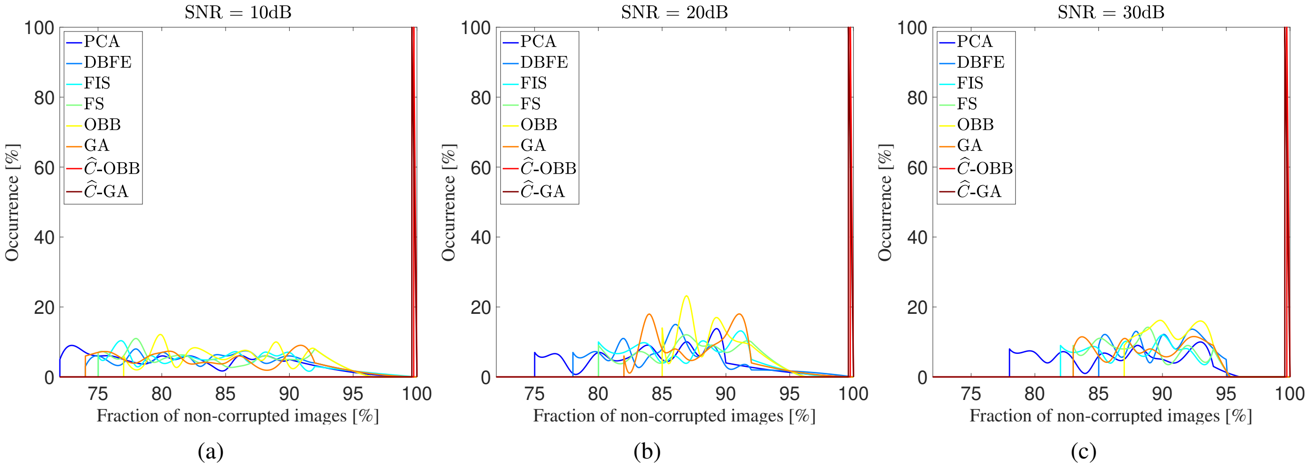
Fig. 7: Overall accuracy achieved by RF classification as a function of the number of selected images by the different schemes on the Beijing dataset. The star markers identify the overall accuracy obtained by RF classification when the selection architectures are left free to determine the best set of images (corresponding to results shown in Table V).

analyzing the Beijing dataset without performing any selection on the records ("No slct." in Table V), as in [48]. The analysis conducted on the whole dataset might be more accurate than the investigation performed when using the image selection algorithms. However, the computational cost associated with this scheme is very high with respect to the reduction delivered by the image selection, as it is clear from Fig. 6. Thus, the search for informative records would save the classifier from investigating non-relevant data, leading to an adequate and effective improvement of the system efficiency. This can be considered as a measure of how the proposed approach can actually guide the decision in ensemble learning and ensemble pruning classification: this effect will be properly explored in future works.

In order to understand how the choice of the images to be processed can actually impact the overall accuracy of the classification, we have conducted additional tests on the image selection schemes outcomes. Specifically, we have forced the five considered image selection architectures to extract the most relevant images when their number was set to $\{q \cdot 5\}_{q=1,\ldots,10}$. Then, we have computed the overall accuracy achieved by means of the RF algorithm. Fig. 7 reports the results. The search for the most informative images can be strongly affected by the number of images that are selected. Indeed, the curves in Fig. 7 show that the accuracy increases as the number of images climbs up to the optimal amount identified by the selection architectures. The accuracy reaches a weak steady behavior as the number of images keeps increasing after this point when selection algorithms based on geometrical and statistical criteria are employed (i.e., PCA, DBFE, FIS). On the contrary, the accuracy degrades when the number of images passes the optimal selection point defined by the schemes in Section III. This effect shows how the search for most relevant images can be very complicated and possibly mislead by improper set-up of the optimization protocols. Moreover, Fig. 7 shows that, although, FIS and $\widehat{\mathcal{C}}$-GA have the same optimal number of images, our approach achieves a better accuracy which emphasizes the precision of our method at selecting the relevant attributes.

From the previous results, it is possible to notice how the ability to describe the considered multimodal scene depends on the ability to extract relevant features from the dataset. It is thus important to assess the performance of the different methods to actually pick informative attributes from the set of records to be processed. To this aim, we conducted additional tests on enriched versions of the Beijing dataset. Specifically, we created 100 datasets where each pixel was characterized by the 53 attributes (corresponding to the $M = 53$ images) that were originally collected in the Beijing dataset, plus other 53 fake attributes that were randomly generated by a Gaussian noise having a variance such that the overall SNR for each pixel would be equal to $\{10, 20, 30\}$ dB. According to the notation we used throughout the paper, for every value of SNR we had then 100 datasets of $M = 106$ images (53 of them being fake images) of 3265×2448 pixels. We used the algorithms that were previously introduced in Section III to select the significant features.

In this case, our attention was focused on understanding whether each method was able to identify the actual attributes that were originally stored in the Beijing dataset and to discriminate them from the fake attributes we added. Hence, we computed for each SNR the fraction of non-corrupted images that were selected by every method, and calculated the occurrence of this quantity over all the datasets we created. Fig. 8 summarizes these results for the SNR=$\{10, 20, 30\}$ dB cases ((a)-(c), respectively). It it possible to appreciate how the proposed $\widehat{\mathcal{C}}$-OBB and $\widehat{\mathcal{C}}$-GA methods are actually able to discriminate the real images in the Beijing dataset from the noisy features we added. The performance of each other algorithm is suffering of a remarkable percentage of wrong selection of features, that can reach up to 25 % of the total feature selection in some cases. This result emphasizes how



(a) (b) (c)

Fig. 8: Distribution of the fraction of non-corrupted images that are selected by the algorithms in Section IV-D when the synthetic dataset obtained by corrupting the Beijing dataset with noise at SNR set to 10dB, 20dB and 30dB ((a) to (c), respectively) is considered. We note that our two methods $\widehat{C}$-OBB and $\widehat{C}$-GA only selected relevant images, contrary to other methods, which selected a significant percentage of fake images.

the proposed approach is able to identify the most informative features in each dataset, and quantifies the gap between the proposed architectures and the algorithms in technical literature in retrieving relevant features from the considered scenes.

2) Pavia dataset: Let us now consider the analysis on the Pavia dataset. In this problem, we aim at classifying the 2800×2800 pixels in the scene in five different classes of air quality (i.e., bad, poor, fair, good, excellent) according to the regulations of the local environmental protection agency. Hence, we are able to derive the confusion matrices for every combination of image selection and classifier. This allows to obtain a more precise characterization of the actual benefits and possible drawbacks of all the methods we have taken into account. Specifically, we can take a look to the confusion matrices that are generated by RF and SVM on the image sets determined by the aforementioned selection algorithms. Tables VI to XIII report these results: they summarize the overall accuracy obtained, as well as the producer's accuracy (i.e., the probability that a value in a given ground truth class was classified correctly) and users's accuracy (i.e., the probability that a predicted value actually represents that class in the ground truth) achieved for every air quality class. For a given class, the producer's accuracy, which reflects how well an area has been classified, is measured as the fraction of correctly predicted values to the total number of ground truth values. While the users's accuracy, which quantifies the reliability of classification, is calculated as the fraction of correctly predicted values to the total number of classified values [59].

From these Tables, it is possible to appreciate how the proposed approach is able to strongly outperform the other methods. In particular, the results obtained for each class display how the information theory-based investigation introduced in this paper is able to provide a remarkable improvement in characterizing the scenes and their specific components. Indeed, from this analysis on an heterogeneous dataset, it is possible to notice that the choice of the classification algorithm does not substantially bias the outcome. Instead, the choice of the selection process is crucial to drive the exploration and reach a solid understanding of the image contents.

In fact, an inadequate reduction of records according to geometrical or statistical properties apparently can jeopardize the investigation of complex datasets (see for instance [48], [60] and [6]). Moreover, such reduction might lead to a strong information loss with respect to the analysis conducted without any record selection (see Table XIV), although the efficiency of the system could be enhanced (see Fig. 6). This aspect becomes particularly evident for the Pavia dataset (more than in the results of the test on Beijing in Table V), as the records are composed of remote sensing and clinical data. Hence, we can state that the proposed approach identifies a valid solution when investigating the significance of records contained in large scale and heterogeneous datasets, such that the accuracy of the analysis can be largely improved with an acceptable precision-efficiency trade-off.

3) IEEE GRSS Data Fusion Contest 2018: Let us finally focus on the IEEE GRSS Data Fusion Contest 2018 [61] dataset, which ground truth is shown in Fig. 9. In this case, in order to highlight the ability of the proposed approach to work at different levels of image fusion, we selected the most relevant bands within the hyperspectral and Lidar records. We hence worked on the whole datacube and run the selection on all the records associated with each pixel. Then, we ran classification based on RF and SVM algorithms.

Figures 10 and 11 show the classification maps obtained on this dataset. Moreover, Table XV reports the overall accuracy and kappa statistics results. In particular, the kappa coefficient measures the agreement between classification and truth values (taking into account the probability of correct classification by chance as well), and it is computed as $K = \frac{N\sum_{i=1}^{R} A_{ii} - \sum_{i=1}^{R} B_i C_i}{N^2 - \sum_{i=1}^{R} B_i C_i} \in [-1, 1]$, where N is the total number of classified values compared to truth values; A_{ii} is the number of values belonging to the truth class i that

TABLE VI: Confusion matrix obtained on Pavia dataset using 16 images selected by PCA out of 35. Class counts ($\times 10^3$) achieved by using RF classification are reported: results obtained by means of SVM analysis are displayed in brackets. Classes from ground truth and classification are reported in columns and rows, respectively. Producer accuracy (PA), users's accuracy (UA) and overall accuracy (OA) results are also reported. For each index and each class, the best result that has been achieved through the results in Tables VI to XIV are shown in red, whilst the second best are shown in blue.

Air quality class	**Bad**	**Poor**	**Fair**	**Good**	**Excellent**	**UA [%]**
Bad	755.3 (764.2)	187.4 (165.1)	110.9 (99.2)	101.1 (100.5)	66.5 (58.1)	61.8 (64.3)
Poor	182.2 (179.5)	408.1 (413.9)	182.3 (184,5)	213.8 (212.2)	82.4 (80)	38.1 (38.6)
Fair	194.3 (193)	113.5 (122.7)	551.2 (564.5)	512.5 (498.1)	102.3 (96.1)	37.3 (38.2)
Good	178 (176)	42.6 (38.4)	72.6 (79.9)	1410.4 (1432.4)	600.2 (588.8)	61.2 (61.8)
Excellent	101.4 (98.5)	32.4 (43.9)	23.8 (12.7)	506.2 (500.8)	1108.6 (1137)	62.5 (63.4)
PA [%]	53.5 (54.1)	52.05 (52.8)	58.6 (60)	51.4 (52.2)	56.5 (58)	**OA [%]** 54 (55)

TABLE VII: Confusion matrix obtained on Pavia dataset using 20 images selected by DBFE out of 35. The same notation of Table VI applies here.

Air quality class	**Bad**	**Poor**	**Fair**	**Good**	**Excellent**	**UA [%]**
Bad	778.1 (787.8)	180.7 (158.7)	88.3 (86.2)	54.4 (59.9)	33.3 (45.2)	68.5 (69.2)
Poor	189.9 (185.3)	420.7 (422)	202.1 (201.4)	113.2 (102.2)	40 (40)	43.5 (44.3)
Fair	180.8 (177)	110.6 (122.7)	553 (566)	809.2 (665.2)	52.3 (56.1)	32.4 (35.6)
Good	172.2 (170.1)	40.8 (42.5)	75.2 (66.7)	1420.4 (1446.1)	710.7 (678.4)	58.7 (60.1)
Excellent	90.2 (91)	31.2 (38.1)	22.2 (20.5)	346.8 (470.6)	1123.7 (1140.3)	69.6 (64.7)
PA [%]	55.1 (55.8)	53.6 (53.8)	58.7 (60.1)	51.7 (52.7)	57.3 (58.1)	**OA [%]** 54.7 (55.6)

TABLE VIII: Confusion matrix obtained on Pavia dataset using 17 images selected by FIS out of 35. The same notation of Table VI applies here.

Air quality class	**Bad**	**Poor**	**Fair**	**Good**	**Excellent**	**UA [%]**
Bad	1002.9 (1040.3)	120.2 (117.9)	42.4 (42.1)	55.2 (60.5)	36.7 (50)	79.7 (79.3)
Poor	290.5 (296.1)	548.8 (555.1)	203.3 (202.9)	100.4 (112.2)	71.7 (68.3)	45.1 (44.9)
Fair	80.4 (48)	55.6 (54.2)	650.8 (651.9)	444.3 (403.6)	123.2 (104.7)	48 (51.6)
Good	20.3 (18.4)	39.3 (36.1)	38.2 (37.7)	1887.8 (1898.8)	348.6 (355.2)	80.8 (80.9)
Excellent	17.1 (8.4)	20.1 (20.7)	6.1 (6.2)	256.3 (268.9)	1379.8 (1381.8)	82.1 (81.9)
PA [%]	71 (73.7)	70 (70.8)	69.2 (69.3)	68.8 (69.2)	70.4 (70.5)	**OA [%]** 69.7 (70.5)

TABLE IX: Confusion matrix obtained on Pavia dataset using 15 images selected by FS out of 35. The same notation of Table VI applies here.

Air quality class	**Bad**	**Poor**	**Fair**	**Good**	**Excellent**	**UA [%]**
Bad	939.8 (906)	109.3 (111.8)	50.3 (59.8)	150.8 (176.6)	90.1 (87.9)	70.1 (67.5)
Poor	300.1 (285.5)	535.5 (527.6)	202.1 (195)	239 (212.3)	100.2 (155.8)	38.9 (38.3)
Fair	115.2 (123.7)	54.4 (66.2)	589.8 (568.2)	350.4 (372.1)	198.6 (218.1)	45 (42.1)
Good	36 (54.1)	43.7 (45.3)	86.5 (90.5)	1720.5 (1679.3)	244.2 (218.3)	80.7 (80.4)
Excellent	20.1 (41.9)	41.1 (33.1)	12.1 (27.3)	283.3 (303.7)	1326.9 (1279.9)	78.8 (75.9)
PA [%]	66.6 (64.2)	68.3 (67.3)	62.7 (60.4)	61.9 (61.2)	67.7 (65.3)	**OA [%]** 65.2 (63.2)

have also been classified as class i; B_i is the total number of truth values belonging to class i; C_i is the total number of predicted values belonging to class i. A maximum value of kappa statistics, $K = 1$, reflects perfect agreement between ground truth and classification, while 0 value corresponds to a random agreement and negative values to no agreement. It is prominent to recall that reaching the best accuracy classification results (as targeted by the contest scenario [61]) is out of the scope of this paper. Nonetheless, the results we obtained are instrumental to show the importance of a reliability and significance assessment of the data to be processed. In fact, it is possible to notice how the use of the proposed approach for record selection can actually enhance the classification results that can be achieved by means of classic schemes like RF and

TABLE X: Confusion matrix obtained on Pavia dataset using 18 images selected by OBB out of 35. The same notation of Table VI applies here.

Air quality class	**Bad**	**Poor**	**Fair**	**Good**	**Excellent**	**UA [%]**
Bad	1020.3 (1035.8)	106.7 (92.3)	42.1 (33)	33.5 (6)	46.7 (65.6)	81.6 (84)
Poor	211.2 (190.3)	540.1 (595.8)	134.3 (102.3)	117.3 (86.3)	80.7 (118.8)	49.8 (21.2)
Fair	86.7 (81.1)	72.3 (54.1)	607.7 (611.5)	400.2 (410.3)	243.5 (140.8)	43 (47.1)
Good	65.8 (52.3)	41.7 (25.6)	130.1 (145.4)	1655.9 (1811)	250.4 (282.4)	77.2 (78.1)
Excellent	27.2 (51.7)	23.2 (16.2)	26.6 (48.6)	537.1 (430.4)	1338.7 (1352.4)	68.5 (71.2)
PA [%]	72.3 (73.4)	74.5 (76)	64.6 (65)	60.2 (66)	68.3 (69)	**OA [%]** 65.8 (68.9)

TABLE XI: Confusion matrix obtained on Pavia dataset using 15 images selected by GA out of 35. The same notation of Table VI applies here.

Air quality class	**Bad**	**Poor**	**Fair**	**Good**	**Excellent**	**UA [%]**
Bad	1030.2 (1059.8)	105.3 (96.3)	39.5 (28.3)	56.4 (45.5)	72.3 (61.4)	79 (82.1)
Poor	197.3 (166.2)	581.7 (588.8)	149.9 (152.8)	100.7 (124)	159.5 (99.9)	48.9 (52)
Fair	96.1 (88.4)	59.5 (74.2)	651 (639.7)	340.2 (256.3)	174.3 (168.2)	49.2 (52.1)
Good	57.3 (69)	32 (16.3)	53.2 (85.8)	1767.1 (1830.2)	287.8 (252.6)	80.4 (78.4)
Excellent	30.3 (27.8)	5.5 (8.4)	47.2 (34.2)	479.6 (488)	1266.1 (1377.9)	69.2 (71.1)
PA [%]	73 (75.1)	74.2 (75.1)	69.2 (68)	64.4 (66.7)	64.6 (70.3)	**OA [%]** 67.5 (70.1)

TABLE XII: Confusion matrix obtained on Pavia dataset using 15 images selected by $\widehat{\mathcal{C}}$-OBB out of 35. The same notation of Table VI applies here.

Air quality class	**Bad**	**Poor**	**Fair**	**Good**	**Excellent**	**UA [%]**
Bad	1303.9 (1322.3)	25.2 (24.8)	9.8 (8.7)	21.6 (22.3)	10.7 (8.3)	**95** (**95.3**)
Poor	65.3 (63.1)	712.6 (714.2)	45.6 (45.1)	48.3 (44.4)	23 (22.7)	**79.6** (**80.3**)
Fair	24.4 (16.3)	23.4 (21.9)	845.7 (846.7)	57.6 (54.2)	34.2 (32.3)	**85.8** (**87.1**)
Good	10.3 (6.5)	16.7 (15.2)	30.3 (29.8)	2532.7 (2538.2)	69.3 (66.1)	**95.2** (**95.5**)
Excellent	7.3 (3)	6.1 (7.9)	9.4 (10.5)	83.8 (84.9)	1822.8 (1830.6)	**94.4** (**94.5**)
PA [%]	**92.4** (**93.7**)	**90.9** (**91.1**)	**89.9** (**90**)	**92.3** (**92.5**)	**93** (**93.4**)	**OA [%]** **92** (**92.5**)

TABLE XIII: Confusion matrix obtained on Pavia dataset using 14 images selected by $\widehat{\mathcal{C}}$-GA out of 35. The same notation of Table VI applies here.

Air quality class	**Bad**	**Poor**	**Fair**	**Good**	**Excellent**	**UA [%]**
Bad	1315.2 (1327.9)	22.5 (21.3)	14.6 (13.9)	21.8 (19.7)	12.4 (13.5)	**94.8** (**95.1**)
Poor	52.3 (46.5)	719.7 (722)	26.6 (25.8)	42.9 (40.3)	20.9 (21.4)	**83.4** (**84.3**)
Fair	29.2 (24.4)	22.6 (20.8)	865.5 (867.4)	53.2 (51)	35.3 (33.5)	**86** (**87**)
Good	11.7 (10)	13.4 (12.9)	24.3 (22.2)	2535.4 (2540.9)	72.6 (68.8)	**95.4** (**95.7**)
Excellent	2.8 (2.4)	5.8 (7)	9.8 (11.5)	90.7 (92.1)	1818.8 (1822.8)	**94.3** (**94.1**)
PA [%]	**93.2** (**94.1**)	**91.8** (**92.1**)	**92** (**92.2**)	**92.4** (**92.6**)	**92.8** (**93**)	**OA [%]** **92.5** (**92.8**)

SVM, and reach performance that are comparable to those obtained by more complex and detailed schemes as in [61].

Indeed, considering the aforesaid results, we can state that data reduction could actually improve the characterization of this multimodal dataset, provided that the data reduction is done following relevant approaches. Further, among the selection algorithms we have considered, the proposed capacity-driven approach is able to distinctly outperform the other selection methods based on eigenvalue analysis, separability enhancement and Fisher information maximization. This effect is even more emphasized by Fig. 12, where the occurrence distribution of the producer accuracies for each class in the dataset is displayed for all the methods in Table XV and Fig. 10 and 11. It is worth recalling that the distribution of producer's accuracy reports information on the precision of the given architecture, i.e., on its ability to classify with low variability. Thus, having a distribution concentrated on the right side of the graph would identify a strong precision of

TABLE XIV: Confusion matrix obtained on Pavia dataset using 35 images when image selection is not performed. The same notation of Table VI applies here.

Air quality class	Bad	Poor	Fair	Good	Excellent	UA [%]
Bad	1032.9 (1031.5)	98.2 (89.9)	35.2 (36.9)	75.6 (68.9)	24.7 (47.4)	81.5 (80.9)
Poor	210.6 (231.5)	537 (518.2)	78.8 (86.5)	183.2 (153.4)	82.4 (77.1)	49.1 (48.5)
Fair	100.3 (92.4)	80.8 (85.3)	660.4 (661.3)	373.8 (366.3)	155.7 (161.1)	48.1 (48.3)
Good	54.8 (36.6)	57.8 (52.3)	106.1 (111.4)	1876.8 (1901.5)	323.3 (302.4)	77.5 (79.1)
Excellent	12.6 (19.2)	10.2 (38.3)	60.3 (44.7)	234.6 (253.9)	1373.9 (1372)	81.2 (79.4)
PA [%]	73.2 (73.1)	68.5 (66.1)	70.1 (70.3)	68.4 (69.3)	70.1 (70)	**OA [%]** 69.9 (69.9)

TABLE XV: Accuracy analysis - IEEE GRSS Data Fusion Contest 2018 data: Overall accuracy (OA) and Kappa statistics (K). For each index, the best result is shown in red, whilst the second is shown in blue. For clarity, the values of kappa statistics were multiplied by 100. M_* represents the optimal number of selected attributes by each method out of 51.

Method	M_*	Classifier	OA [%]	K ×100
PCA	34	RF	51.5	50.9
		SVM	50.7	50.3
DBFE	37	RF	49.7	49.5
		SVM	49.6	47.8
FIS	35	RF	60.2	60
		SVM	59.9	59.3
FS	**32**	RF	53.4	54.2
		SVM	53.1	53.9
OBB	33	RF	57.8	56.6
		SVM	58.1	58
GA	**32**	RF	57.2	55.8
		SVM	53.8	57.3
$\widehat{\mathcal{C}}$**-OBB**	**31**	RF	**76**	**75.9**
		SVM	74.9	73.8
$\widehat{\mathcal{C}}$**-GA**	**32**	RF	**77.5**	**76.9**
		SVM	75.7	74.3
No select.	51	RF	40	39.8
		SVM	39.7	39.6

the given framework in assigning labels on the given dataset. As for Fig. 12, it is possible to observe how the producer's accuracy trend shifts to higher values when $\widehat{\mathcal{C}}$-OBB and $\widehat{\mathcal{C}}$-GA are employed (as expected in an ideal classification), whilst for the other methods the majority of the producer's accuracies are concentrated around values in the interval 10-30 %. Nonetheless, it is worth noticing that this performance improvement translates in a higher computational cost, as reported in Fig. 6, which shows how the execution time of the schemes basically follows the trends already commented for the Beijing and Pavia datasets.

V. Conclusion

In this paper, the limits of multimodal remote sensing data analysis in terms of the maximum reliability and accuracy under optimal conditions are investigated. Our approach leads to the definition of two indices that can be easily computed before the actual processing would take place. Thus, the main contributions of this paper are:

- The introduction of an information-theory based approach to characterize the maximum information extraction performance as a function of data statistics and parameters that characterize any multimodal scenario;
- The definition of two indices that quantify the maximum accuracy and reliability of the analysis that can take place over the considered multimodal records;
- Asymptotic approximations of the aforementioned indices, so that these quantities can be estimated during preprocessing, avoiding the use of difficult numerical integration and root finding techniques;
- Large experimental validation that shows the consistency of our approach, as well as the validity of the proposed indices.

The experimental results show that image selection performed by means of schemes that are not properly addressing informativity maximization might prevent the data analysis scheme from reaching the optimal interpretation of the records, and in some cases might even degrade the investigation with respect to a scenario where no data reduction is applied. This emphasizes the urgency of a proper assessment of reliability in remote sensing data analysis, and especially in multimodal remote sensing investigation, so that a precise understanding and interpretation of the Earth's surface phenomena can be retrieved.

For sake of clarity, we have introduced the reliability and capacity metrics for decision-level fusion. Future works will be devoted to the extension of the results to feature-level fusion. Nevertheless, the development and implementation of platforms that are able to perform enhanced multimodal remote sensing data analysis by considering the proposed metrics during preprocessing for automatic image selection have been presented in this paper. Large scale efficient dimensionality reduction, and near real-time information extraction will be the target of further works.

To accurately assess the capacity and reliability of highly non-linear approaches of data fusion, such as the ones based on deep learning, a further improvement of the proposed model could be examined by considering spatially correlated noises and non-linear imperfections. Moreover, multimodal remote sensing analysis is subject to many mismatches due to the heterogeneity of data, such as an improper co-registration or

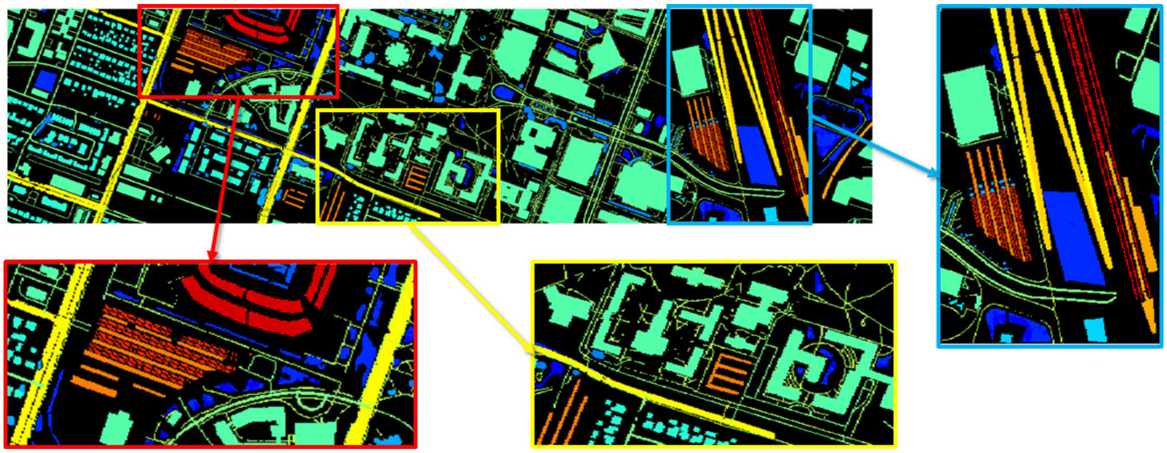


Fig. 9: Ground truth of the 20 classes identified in the IEEE GRSS Data Fusion Contest 2018.

alignment. An interesting perspective of this work would be to investigate how such mismatches could affect the capacity of the analysis to extract information.

Furthermore, by following the approach we proposed in this paper, it would be possible to assess the actual importance of different modalities in the characterization of specific regions, as a result of the impact of different illumination conditions, registration conditions, and preprocessing operations. Thus, it is possible to expect that this approach might improve the transfer learning performance on multimodal datasets by producing an adaptive scheme to weigh the significance of the different inputs in different areas, and hence to enhance understanding the physical phenomena.

Appendix A
Computation of the Capacity

The capacity of the system in (1) is defined as the mutual information between $\underline{\underline{\widehat{Y}}}$ and $\underline{\underline{\widehat{X}}}$, maximized over the input distribution,

$$\begin{aligned} \mathcal{C} &= \max_{p(\underline{\underline{X}})} I(\underline{\underline{\widehat{X}}};\underline{\underline{\widehat{Y}}}) = \max_{p(\underline{\underline{X}})} \sum_{l=1}^{L} I(\underline{\hat{x}}_l;\underline{\hat{y}}_l(\underline{\underline{H}}_l)), \qquad (16)\\ &= \max_{p(\underline{\underline{X}})} \sum_{l=1}^{L} \mathbb{E}_{\underline{\underline{H}}_l}\left[I(\underline{\hat{x}}_l;\underline{\hat{y}}_l|\underline{\underline{H}}_l)\right]. \qquad (17)\end{aligned}$$

The distribution of the input $\underline{\hat{x}}_l$ that maximizes $\mathcal{C}$ is Gaussian with covariance $\mathbb{E}[\underline{\hat{x}}_l\underline{\hat{x}}_l^T]$ ($\underline{\hat{x}}_l$ is assumed to be zero-mean) [13], [14]. Using the Gaussian assumptions [62], it follows that

$$\mathcal{C} = \max_{\hat{P}_l} \sum_{l=1}^{L} \mathbb{E}_{\underline{\underline{H}}_l}\left[\log_Q \left| \underline{\underline{I}}_R + \frac{\underline{\underline{H}}_l \mathbb{E}[\underline{\hat{x}}_l\underline{\hat{x}}_l^T]\underline{\underline{H}}_l^T}{N_0}\right|\right], \qquad (18)$$

where $\hat{P}_l = \mathbb{E}[\underline{\hat{x}}_l^T\underline{\hat{x}}_l]$, $\log_Q x = \frac{\ln x}{\ln Q}$ and Q refers to the quantization levels of $\underline{\hat{x}}_l$, and $|.|$ is the determinant operator. We can simplify the expression of $\mathcal{C}$ as follows,

$$\sum_{l=1}^{L} \log_Q \left| \underline{\underline{I}}_R + \frac{\underline{\underline{H}}_l \mathbb{E}[\underline{\hat{x}}_l\underline{\hat{x}}_l^T]\underline{\underline{H}}_l^T}{N_0}\right| \qquad (19)$$

$$\begin{aligned} &= \sum_{l=1}^{L} \log_Q \left| \underline{\underline{I}}_R + \frac{\underline{\underline{\Lambda}}_l \mathbb{E}[\underline{\tilde{x}}_l\underline{\tilde{x}}_l^T]\underline{\underline{\Lambda}}_l^T}{N_0}\right| \qquad (20)\\ &\leq \sum_{l=1}^{L} \log_Q \prod_{r=1}^{R}\left(1+\frac{\lambda_{lr}\hat{P}_{lr}}{N_0}\right) \qquad (21)\\ &= \sum_{l=1}^{L}\sum_{r=1}^{R} \log_Q \left(1+\frac{\lambda_{lr}\hat{P}_{lr}}{N_0}\right) \qquad (22)\end{aligned}$$

where (20) follows using that $\underline{\underline{H}}_l = \underline{\underline{U}}_l\underline{\underline{\Lambda}}_l\underline{\underline{V}}_l^T$, and (21) follows using Hadamard's inequality. Hadamard's inequality states that the determinant of a positive definite matrix is less than or equal to the product of its diagonal elements.

Considering in addition the power constraint $\frac{1}{LR}\sum_{l=1}^{L}\sum_{r=1}^{R}\hat{P}_{lr} = \hat{P}$, we define the capacity as in [31], as follows

$$\mathcal{C} = \max_{\substack{\hat{P}_{lr} \\ \text{s.t.} \sum_{l=1}^{L}\hat{P}_l=\hat{P}}} \sum_{l=1}^{L}\sum_{r=1}^{R} \mathbb{E}_{\zeta_{lr}}\left[\log_Q\left(1+\frac{\zeta_{lr}\hat{P}_{lr}}{\hat{P}}\right)\right]. \qquad (23)$$

where $\zeta_{lr} = \frac{\lambda_{lr}\hat{P}_{lr}}{N_0}$.

Appendix B
Uniqueness of Z_0

Further details can be found in [14], from where the steps of the following proof is inspired. Using the definition of $p_\zeta(\zeta)$ in (8), (7) becomes,

$$\sum_{r=1}^{R}\kappa_r\int_{\nu}^{\infty}\left(\frac{1}{\nu}-\frac{1}{\zeta}\right)\psi_\zeta\left[\mathcal{L}_{r-1}^{R(M-1)}(\zeta)\right]^2 \mathrm{d}\zeta = RL\bar{\zeta}, \qquad (24)$$

where $\psi_\zeta = e^{-\zeta}\zeta^{R(M-1)}$. Let us define $\phi_r(\zeta,\nu) = \left(\frac{1}{\nu}-\frac{1}{\zeta}\right)e^{-\zeta}\zeta^{R(M-1)}\left[\mathcal{L}_{r-1}^{R(M-1)}(\zeta)\right]^2$ [14]. Further, let us consider

$$\Phi(\nu) = \sum_{r=1}^{R}\kappa_r\int_{\nu}^{\infty}\phi_r(\zeta,\nu)\mathrm{d}\zeta - RL\bar{\zeta}. \qquad (25)$$

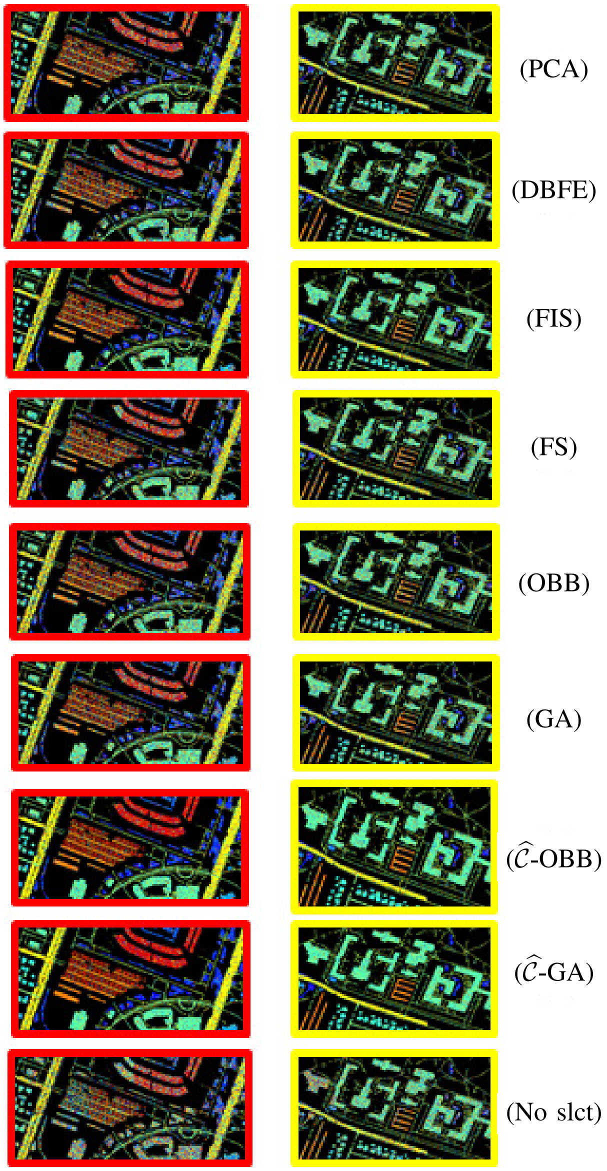


Fig. 10: Details of the classification results on the areas in the red and yellow boxes in Fig. 9. Classification results are achieved by using RF algorithm on the IEEE GRSS Data Fusion Contest 2018 dataset when image selection is performed by means of PCA, DBFE, FIS, FS, OBB, GA, $\widehat{\mathcal{C}}$-OBB and $\widehat{\mathcal{C}}$-GA, and when no image selection is performed. The maps show that our approaches ensure a better classification, especially fo the stadium seats in red, and the roads class in yellow.

Then, it is possible to prove that $\partial\Phi(\nu)/\partial\nu < 0$ and $\partial^2\Phi(\nu)/\partial\nu^2 > 0$ for $\nu > 0$. Moreover, thanks to the normalization property of the probability distribution function, the following holds,

$$\lim_{\nu\to 0^+} \int_\nu^\infty e^{-\xi}\xi^{R(M-1)} \left[\mathcal{L}_{r-1}^{R(M-1)}(\xi)\right]^2 \mathrm{d}\xi = \kappa_r^{-1}. \quad (26)$$

Analogously, we can write as follows, considering

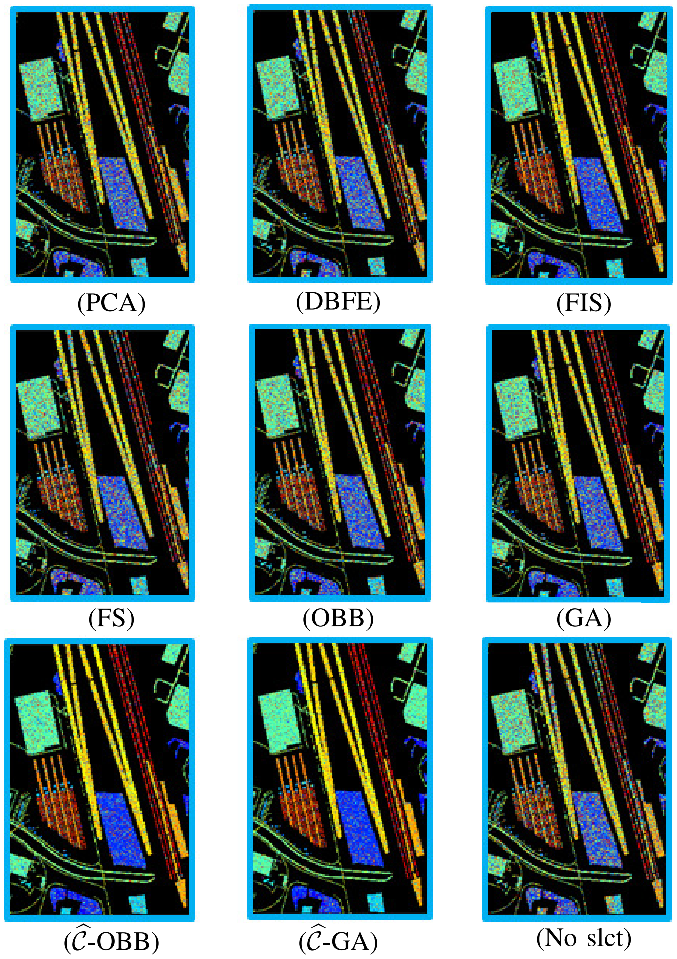


Fig. 11: Details of the classification results on the areas in the light blue box in Fig. 9. Classification results are achieved by using RF algorithm on the IEEE GRSS Data Fusion Contest 2018 dataset when image selection is performed by means of PCA, DBFE, FIS, FS, OBB, GA, $\widehat{\mathcal{C}}$-OBB and $\widehat{\mathcal{C}}$-GA, and when no image selection is performed. Our approaches produce better results, especially fo the non-residential buildings in green and the railways in yellow.

$$\mathcal{U}(\nu) = \int_\nu^\infty e^{-\xi}\xi^{R(M-1)-1} \left[\mathcal{L}_{r-1}^{R(M-1)}(\xi)\right]^2 \mathrm{d}\xi,$$

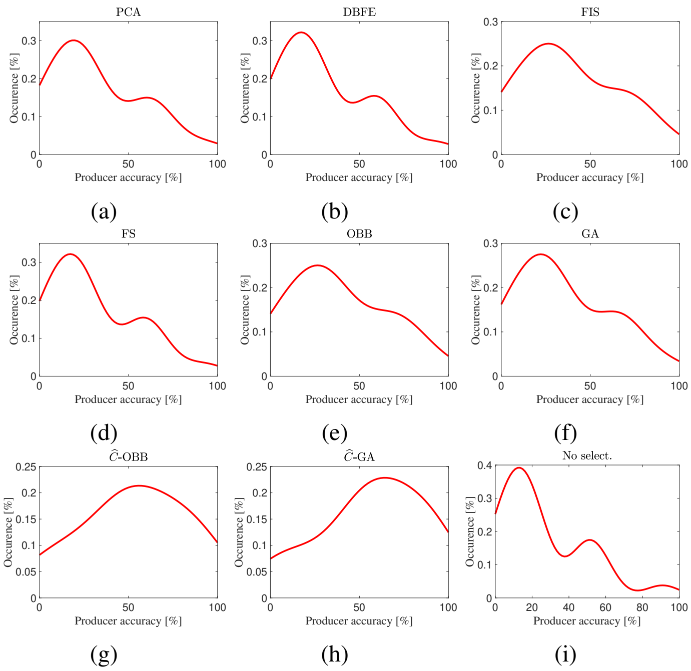
$$\lim_{\nu\to 0^+} \mathcal{U}(\nu) = \frac{\partial^{r-1}}{\partial w^{r-1}} S(R(M-1),w)\bigg|_{w=0} T(R,M,r), \quad (27)$$

where

$$T(R,M,r) = \frac{\Gamma(R(M-1)\Gamma(R(M-1)+r)}{\Gamma(R(M-1)+1)((r-1)!)^2}, \quad (28)$$

being $\Gamma(t) = \int_0^\infty u^{t-1}e^{-t}\mathrm{d}u$. Moreover, $S(R(M-1),w) = F[R(M-1)/2; (R(M-1)+1)/2; R(M-1)+1; 4w/(1+w)^2]/[(1-w)(1+w)^{R(M-1)}]$, where

$$F[t_1;t_2;t_3;t_4] = \sum_{j=0}^{\infty} \frac{(t_1)_j (t_2)_j}{(t_3)_j} \frac{t_4^j}{j!}, \quad (29)$$



(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

Fig. 12: Producer's accuracy distributions for classification results achieved by using RF algorithm on the IEEE GRSS Data Fusion Contest 2018 dataset when image selection is performed by means of PCA, DBFE, FIS, FS, OBB, GA, $\widehat{\mathcal{C}}$-OBB and $\widehat{\mathcal{C}}$-GA ((a) to (h), respectively), and when no image selection is performed (i). In contrast to our approaches (g) and (h), the other methods show a similar trend with a high concentration towards small values of the producer's accuracy, which reflects their instability.

where we have $(v)_j = \prod_{h=0}^{j-1} v + h$, and $(v)_0 = 1$. Thus, by means of an hypergeometric transformation to be applied on (27) according to the aforementioned definitions,

$$\lim_{\nu \to 0^+} \mathcal{U}(\nu) = \frac{(R(M-1)+r-1)!}{(R(M-1))(r-1)!}. \tag{30}$$

Therefore, using this result and (26) in the definition of $\Phi(\nu)$, $\lim_{\nu\to 0^+} \Phi(\nu) = +\infty$, and $\lim_{\nu\to\infty} \Phi(\nu) = -RL\bar{\zeta}$.

Thus, since $\partial\Phi(\nu)/\partial\nu < 0$, $\Phi(\nu)$ has a unique value s.t. $\Phi(\nu) = 0$. Hence, for any $\bar{\zeta} > 0$ there is a unique Z_0 that satisfies (7).

Appendix C
Details of Z_0 computation

The term $\mathcal{L}_s^t(u)$ in (9) can be written, using the binomial expansion, as follows:

$$\mathcal{L}_s^t(u) = \sum_{j=0}^{s} (-1)^j \binom{s+t}{s-j} u^j / j!. \tag{31}$$

With this in mind, (24) becomes,

$$L \sum_{r=1}^{R} \kappa_r \sum_{j_1=0}^{r-1} \sum_{j_2=0}^{r-1} \frac{(-1)^{j_1+j_2}}{j_1! j_2!} J_1 J_2 \mathcal{G}_{j_1 j_2}(\nu) = RL\bar{\zeta}, \tag{32}$$

where $\mathcal{G}_{j_1j_2}(\nu) = \int_\nu^\infty \left(\frac{1}{\nu} - \frac{1}{\zeta}\right) e^{-\zeta}\zeta^{R(M-1)+j_1+j_2}\mathrm{d}\zeta$, as $j_1 + j_2 = 0, 1, \ldots, 2(R-1)$. Further, since $\int_\nu^\infty e^{-\zeta}\zeta^{RM}\mathrm{d}\zeta = (RM)!e^{-\nu}\sum_{k=0}^{RM}\nu^k/k!$, we can write $\mathcal{G}_{j_1j_2}(\nu) = \Gamma(\beta_1,\nu)/\nu - \Gamma(\beta_2,\nu)$, so (32) can be written as (10).

Acknowledgment

The authors would like to thank the National Center for Airborne Laser Mapping and the Hyperspectral Image Analysis Laboratory at the University of Houston for acquiring and providing the data used in this study, and the IEEE GRSS Image Analysis and Data Fusion Technical Committee.

This work is funded in part by Centre for Integrated Remote Sensing and Forecasting for Arctic Operations (CIRFA) and the Research Council of Norway (RCN Grant no. 237906), and by the Fram Center under the Automised Large-scale Sea Ice Mapping (ALSIM) "Polhavet" flagship project.

References

[1] L. Gomez-Chova, D. Tuia, G. Moser, and G. Camps-Valls. Multimodal classification of remote sensing images: A review and future directions. *Proceedings of the IEEE*, 103(9):1560–1584, Sept. 2015.

[2] A. Voisin, V. A. Krylov, G. Moser, S. B. Serpico, and J. Zerubia. Supervised classification of multisensor and multiresolution remote sensing images with a hierarchical copula-based approach. *IEEE Trans. Geosci. Remote Sens.*, 52(6):3346–3358, Jun. 2014.

[3] S. Chlaily, C. Ren, P.-O. Amblard, O. J. J. Michel, P. Comon, and C. Jutten. Information-estimation relationship in mismatched gaussian channels. *IEEE Signal Processing Letters*, 24(5):688–692, 2017.

[4] F. Sedighin, M. Babaie-Zadeh, B. Rivet, and C. Jutten. Multimodal soft nonnegative matrix co-factorization for convolutive source separation. *IEEE Trans. Signal Processing*, 65(12):3179–3190, 2017.

[5] N. Longbotham et al. Multi-modal change detection, application to the detection of flooded areas: Outcome of the 2009-2010 data fusion contest. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 5(1):331–342, Feb 2012.

[6] M. Datcu, H. Daschiel, A. Pelizzari, M. Quartulli, A. Galoppo, A. Colapicchioni, M. Pastori, K. Seidel, P. G. Marchetti, and S. D'Elia. Information mining in remote sensing image archives: system concepts. *IEEE Transactions on Geoscience and Remote Sensing*, 41(12):2923–2936, Dec 2003.

[7] M. Dalla Mura, S. Prasad, F. Pacifici, P. Gamba, J. Chanussot, and J. A. Benediktsson. Challenges and opportunities of multimodality and data fusion in remote sensing. *Proceedings of the IEEE*, 103(9):1585–1601, Sep. 2015.

[8] P. Ghamisi, B. Rasti, N. Yokoya, Q. Wang, B. Höfle, L. Bruzzone, F. Bovolo, M. Chi, K. Anders, R. Gloaguen, P. Atkinson, and J. Benediktsson. Multisource and multitemporal data fusion in remote sensing: A comprehensive review of the state of the art. *IEEE Geoscience and Remote Sensing Magazine*, 7:6–39, 03 2019.

[9] M. Schmitt and X. X. Zhu. Data fusion and remote sensing: An ever-growing relationship. *IEEE Geoscience and Remote Sensing Magazine*, 4(4):6–23, 2016.

[10] D. Lazer, R. Kennedy, G. King, and A. Vespignani. The parable of google flu: traps in big data analysis. *Science*, 343:1203–1205, March 2014.

[11] D. Mackay. *Information Theory, Inference and Learning Algorithms*. Cambridge Univ. Press, 2003.

[12] J. Zhang et al. Evolutionary computation meets machine learning: A survey. *IEEE Computational Intelligence Magazine*, 6(4):68–75, 2011.

[13] E. Biglieri, R. Calderbank, A. Constantinides, A. Goldsmith, A. Paulraj, and H.V. Poor. *MIMO Wireless communications*. Cambridge Univ. Press, Cambridge, UK, 2007.

[14] M. S. Alouini and A. J. Goldsmith. Capacity of rayleigh fading channels under different adaptive transmission and diversity-combining techniques. *IEEE Trans. Veh. Technol.*, 48:1165–1181, 1999.

[15] A. Marinoni and P. Gamba. An efficient approach for local affinity pattern detection in remotely sensed big data. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 8(10):4622–4633, Oct. 2015.

[16] W. Li, Q. Du, and M. Xiong. Kernel collaborative representation with tikhonov regularization for hyperspectral image classification. *IEEE Geosci. Remote Sens. Lett.*, 12(1):48–52, 2015.

[17] G. Camps-Valls, J. Munoz-Mari, L. Gomez-Chova, K. Richter, and J. Calpe-Maravilla. Biophysical parameter estimation with a semisupervised support vector machine. *IEEE Geoscience and Remote Sensing Letters*, 6(2):248–252, April 2009.

[18] G. Camps Valls and L. Bruzzone. *Kernel Methods for Remote Sensing Data Analysis*. Wiley and Sons Inc, 2009.

[19] J. Chen, J. Xia, P. Du, J. Chanussot, Z. Xue, and X. Xie. Kernel supervised ensemble classifier for the classification of hyperspectral data using few labeled samples. *Remote Sensing*, 8(7):601, 2016.

[20] B. Song, J.Li, M. Dalla Mura, P.Li, A.Plaza, J.M. Bioucas-Dias, J.A. Benediktsson, and J.Chanussot. Remotely sensed image classification using sparse representations of morphological attribute profiles. *IEEE Trans. Geosci. Remote Sens.*, 52(8):5122–5136, 2014.

[21] X. Li, S. Zhang, X. Pan, P. Dale, and R. Cropp. Straight road edge detection from high-resolution remote sensing images based on the ridgelet transform with the revised parallel-beam radon transform. *Int. J. Remote Sens.*, 31:5041–5059, 2010.

[22] G. Camps-Valls, L. Gómez-Chova, J. Muñoz-Marí, J. Vila-Francés, J. Amorós-López, and J. Calpe-Maravilla. Retrieval of oceanic chlorophyll concentration with relevance vector machines. *Remote Sensing of Environment*, 105(1):23 – 33, 2006.

[23] K. Song, D. Lu, L. Li, S. Li, Z. Wang, and J. Du. Remote sensing of chlorophyll-a concentration for drinking water source using genetic algorithms (ga)-partial least square (pls) modeling. *Ecological informatics*, 10:25–36, 2012.

[24] H. Nguyen, K. Koike, and M. Nhuan. Improved accuracy of chlorophyll-a concentration estimates from MODIS imagery using a two-band ratio algorithm and geostatistics: As applied to the monitoring of eutrophication processes over Tien Yen Bay (Northern Vietnam). *Remote Sensing*, 6:421–442, 12 2013.

[25] S. Ihara. On the capacity of channels with additive non-gaussian noise. *Information and Control*, 37(1):34 – 39, 1978.

[26] G. H. Golub and C. F. Van Loan. *Matrix computations*. Baltimore, MD: The Johns Hopkins University Press, 3nd edition, 1996.

[27] G. Blower. *Random Matrices: High Dimensional Phenomena*. London Mathematical Society Lecture Note Series. Cambridge University Press, 2009.

[28] J. Vergara and P.A. Estevez. A review of feature selection methods based on mutual information. *Neural Comput. and Applications*, 24:175–186, 2014.

[29] A. El Gamal and Y.-H. Kim. *Network Information Theory*. Cambridge University Press, 2011.

[30] T.W. Anderson. *An Introduction to Multivariate Statistical Analysis*. Wiley Series in Probability and Statistics. Wiley, 2003.

[31] A. J. Goldsmith and P. P. Varaiya. Capacity of fading channels with channel side information. *IEEE Trans. Inform. Theory*, 43:1986?1992, 1997.

[32] A. Kraskov, H. Stogbauer, and P. Grassberger. Estimating mutual information. *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.*, 69(6), June 2004.

[33] S. Boyd and L. Vandenberghe. *Convex optimization*. New York, NY: Cambridge University Press, 2004.

[34] S. Chakrabarti, J. Judge, T. Bongiovanni, A. Rangarajan, and S. Ranka. Disaggregation of remotely sensed soil moisture in heterogeneous landscapes using holistic structure-based models. *IEEE Transactions on Geoscience and Remote Sensing*, 54(8):4629–4641, 2016.

[35] D. Tuia, R. Flamary, and M. Barlaud. To be or not to be convex? a study on regularization in hyperspectral image classification. In *Proc. of IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, Milan, Italy, July 2015.

[36] K. Svanberg. The method of moving asymptotes - a new method for structural optimization. *Intl. J. for Num. Meth. in Engineering*, 24:359–373, 1987.

[37] A.A. Zhigliavsky and J Pinter. *Theory of global random search*. Dordrecht, the Netherlands: Kluwer Academic Publishers, 1991.

[38] R. Horst and H. Tuy. *Global Optimization - Deterministic Approaches*. Springer, Berlin / Heidelberg / New York, 3rd edition, 1996.

[39] W. Forster. Homotopy methods. In R. Horst and P.M. Pardalos, editors, *Handbook of Global Optimization. Nonconvex Optimization and Its Applications*. Springer, Boston, MA, 1995.

[40] J. Mockus, W. Eddy, A. Mockus, L. Mockus, and G. Reklaitis. *Bayesian Heuristic Approach to Discrete and Global Optimization*. Springer, 1996.

[41] Q. Zheng and D. Zhuang. Integral global minimization: algorithms, implementations and numerical tests. *Journal of Global Optimization*, 7:421–454, 1995.

[42] S. Voss, S. Martello, I.H. Osman, and C. Roucairol. *Meta-Heuristics: Advances and Trends in Local Search Paradigms for Optimization*. Kluwer Academic Publishers, Dordrecht / Boston / London, 1999.

[43] K.A. Dill, A.T. Phillips, and J.B. Rosen. Molecular structure prediction by global optimization. In I.M. Bomze, T. Csendes, R. Horst, and P.M. Pardalos, editors, *Developments in Global Optimization*. Kluwer Academic Publishers, Dordrecht / Boston / London, 1997.

[44] J. J. More' and Z. Wu. Global continuation for distance geometry problems. *SIAM Journal on Optimization*, 7(3):814–836, 1997.

[45] A.V. Levy and S. Gomez. The tunneling method applied for the global minimization of functions. In P.T. Boggs, editor, *Numerical Optimization*. SIAM, Philadelphia, PA, 1985.

[46] J.M. Fowkes, N.I.M. Gould, and C.L. Farmer. A branch and bound algorithm for the global optimization of hessian lipschitz continuous functions. *Journal of Global Optimization*, 56(4):1791–1815, 2013.

[47] 2018 IEEE GRSS Data Fusion Contest. online. http://www.grss-ieee.org/community/technical-committees/data-fusion".

[48] A. Marinoni, G.C. Iannelli, and P. Gamba. An information theory-based scheme for efficient classification of remote sensing data. *IEEE Trans. on Geoscience and Remote Sensing*, 2017.

[49] A. Dagliati, A. Marinoni, C. Cerra, P. Decata, L. Chiovato, P. Gamba, and R. Bellazzi. Integration of administrative, clinical, and environmental data to support the management of type 2 diabetes mellitus: From satellites to clinical care. *Journal of Diabetes Science and Technology*, 10(1):19–26, 2016.

[50] L. Bruzzone. An approach to feature selection and classification of remote sensing images based on the bayes rule for minimum cost. *IEEE Trans. Geosci. Remote Sens.*, 38(1):429–438, 2000.

[51] R.D. Phillips, L.T. Watson, R.H. Wynne, and C.E. Blinn. Feature reduction using a singular value decomposition for the iterative guided spectral class rejection hybrid classifier. *ISPRS Journal of Photogrammetry and Remote Sensing*, 64:107–116, 2009.

[52] C. Lee and D. Landgrebe. Feature extraction based on decision boundaries. *IEEE Trans. Pattern Anal. Machine Intell.*, 15:388–400, 1993.

[53] Q. Gu, Z. Li, and J. Han. Generalized fisher score for feature selection. *Proc. of 27th Conf. on Uncertainty in Artificial Intelligence (UAI11)*, 2012.

[54] A. Zafari, R. Zurita-Milla, and E. Izquierdo-Verdiguier. Evaluating the performance of a random forest kernel for land cover classification. *Remote Sensing*, 11(5), 2019.

[55] J. Xia, N. Falco, J. Benediktsson, P. Du, and J. Chanussot. Hyperspectral image classification with rotation random forest via KPCA. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, PP, 12 2016.

[56] C. Strobl, A.-L. Boulesteix, and T. Augustin. Unbiased split selection for classification trees based on the Gini index. *Computational Statistics and Data Analysis*, 52(1):483–501, 2007.

[57] T.K. Ho. A data complexity analysis of comparative advantages of decision forest constructors. *Pattern Analysis and Applications*, 5(2):102–112, 2002.

[58] C.J.C. Burges. A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, 2:121–167, 1998.

[59] S. V. Stehman. Selecting and interpreting measures of thematic classification accuracy. *Remote Sensing of Environment*, 62(1):77 – 89, 1997.

[60] W. Li and Q. Du. A survey on representation-based classification and detection in hyperspectral remote sensing imagery. *Pattern Recognition Letters*, 2015.

[61] Y. Xu, B. Du, L. Zhang, D. Cerra, M. Pato, E. Carmona, S. Prasad, N. Yokoya, R. Hänsch, and B. Le Saux. Advanced multi-sensor optical remote sensing for urban land use and land cover classification: Outcome of the 2018 IEEE GRSS Data Fusion Contest. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(6):1709–1724, June 2019.

[62] T. Cover and J. Thomas. *Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing)*. Wiley-Interscience, New York, NY, USA, 2006.