# STATISTICAL TESTS FOR HYPERSPECTRAL CODED DATA UNSUPERVISED CLASSIFICATION

Trung-tin Dinh, Hervé Carfantan, Antoine Monmayrant, Simon Lacroix

# STATISTICAL TESTS FOR HYPERSPECTRAL CODED DATA UNSUPERVISED CLASSIFICATION

*Trung-Tin Dinh*[1,2], *Hervé Carfantan*[1], *Antoine Monmayrant*[2], *Simon Lacroix*[2]

[1]: IRAP, 14 Avenue Edouard Belin, 31400 Toulouse, France
[2]: LAAS, 7 Avenue Colonel Roche, 31400 Toulouse, France

## ABSTRACT

We propose a novel method for unsupervised classification of coded hyperspectral acquisitions using a DD-CASSI (*Double Disperser - Coded Aperture Spectral Snapshot Imager*) system, which reduces the number of required acquisitions, typically by an order of magnitude. Leveraging the Separability Assumption (SA) and non-parametric Gaussianity statistical tests, our approach identifies homogeneous regions, which are areas of pixels made of the same material, and determines their unique spectral signatures directly from the coded measurements. By combining these statistical tests with spatial characteristics from panchromatic images, our iterative method effectively classifies regions without reconstructing the entire hyperspectral cube. This approach demonstrates the potential for accurate classification with minimal data, paving the way for optimized hyperspectral data analysis.

***Index Terms***— hyperspectral coded data, unsupervised classification, statistical tests, DD-CASSI

## 1. INTRODUCTION

Conventional hyperspectral (HS) imaging techniques require numerous acquisitions and necessitate spatial or spectral scanning to fill the HS cube $\mathcal{O} \in \mathbb{R}^{R\times C\times W}$, where $(R, C)$ are the spatial dimensions and $W$ is the spectral dimension. In our project, we utilize a controllable coded aperture imager of the DD-CASSI type [1] (Fig. 1) to analyze the hyperspectral scene, typically requiring ten times fewer acquisitions than classical hyperspectral imagers.

The main component of this device is a controllable micro-mirror mask, the DMD (*Digital Micromirror Device*), which acts as a spatio-spectral mask. It allows the selection at each pixel of a combination of spectral bands to be acquired. This snapshot technique reduces both data volume and acquisition times, and increases the signal-to-noise ratio (SNR) at each pixel by combining light from multiple spectral bands. The acquired data are referred to as coded data and one coded

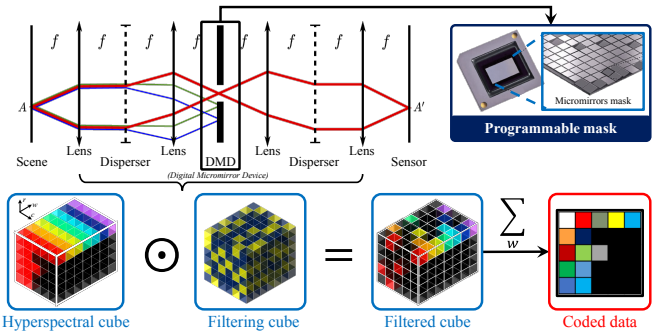


**Fig. 1**. Coded acquisition with DD-CASSI system

acquisition process can be expressed as follow:

$$\mathcal{D} = \sum_{w=1}^{W} \mathcal{H} \odot \mathcal{O} + \mathcal{N}, \tag{1}$$

where $\mathcal{D} \in \mathbb{R}^{R\times C}$ is the coded data, $\odot$ denotes the Hadamard (element-wise) product, $\mathcal{H} \in \mathbb{R}^{R\times C\times W}$ a filtering cube which represents the DMD configuration, and $\mathcal{N}$ the acquisition noise. By vectorization, and stacking for $S$ successive acquisitions, with different DMD configurations, it can be rewritten as:

$$\boldsymbol{d} = \mathbf{H}\boldsymbol{o} + \boldsymbol{n}, \tag{2}$$

where $\boldsymbol{d} \in \mathbb{R}^{RCS\times 1}$ is the coded data, $\mathbf{H} \in \mathbb{R}^{RCS\times RCW}$ is a representative matrix of the instrument and DMD configurations, $\boldsymbol{o} \in \mathbb{R}^{RCW\times 1}$ is the HS cube, and $\boldsymbol{n}$ the corresponding acquisition noise. Our goal is to analyze the HS cube from a small number of acquisitions $S \ll W$.

In the literature, many supervised and unsupervised methods for hyperspectral image classification have been proposed. However, these methods typically require working with the entire hyperspectral cube, either obtained by scanning or reconstructed from compressed data. The study by [2] proposes a supervised classification method using a 3D convolutional neural network for coded hyperspectral data acquired with a DD-CASSI. This method optimizes the configuration of the device's micromirror mask according to the scene, avoiding the need for full hyperspectral cube reconstruction. However, such learning-based methods require a

sufficiently large database to validate the model's accuracy and avoid overfitting, which is not feasible in practice due to the limited number of publicly-accessible labeled hyperspectral databases.

In practice, in most cases, we cannot have *a priori* knowledge about the observed scene, such as spectral signatures or the number of materials. Therefore, supervised classification methods cannot be applied in all scenarios due to the lack of endmembers, and conventional unsupervised classification methods are also not feasible. Moreover, reconstruction can introduce errors affecting the quality of the results, and the volume of reconstructed data is much larger than that of classification results, which are presented as a class map with dimensions $\mathbb{R}^{R\times C}$.

These observations led us to propose an unsupervised classification algorithm directly from coded data, without going through the reconstruction step and without requiring reference spectra or prior knowledge of the number of classes. The objective of this algorithm is to classify homogeneous regions (all pixels from the same material) and emphasize regions requiring additional information.

## 2. UNSUPERVISED CLASSIFICATION ALGORITHM

This algorithm is based on three major computational pillars: the estimation of the reference spectrum in a homogeneous region, the prediction of coded data using a simple model of intraclass spectral variability, and Gaussianity statistical tests.

### 2.1. Estimation of reference spectrum using Separability Assumption (SA)

For a given pixel $n$ of the HS cube, the number of data in $\boldsymbol{d}_n$ $(S)$ is lower than the spectrum length in $\boldsymbol{o}_n$ $(W)$ as $S \ll W$. Therefore, estimating the spectrum from the corresponding coded data is an underdetermined inverse problem that requires prior knowledge.

Our approach to compensate for the lack of information is to account for the separability assumption [3], i.e. that some regions composed of adjacent pixels correspond to a same material and thus share similar spectra, possibly with a different intensity that can be estimated in practice from the panchromatic image. For such a region $r$, the corresponding HS model can be formulated as $\boldsymbol{o}_r = (\boldsymbol{p}_r \boxtimes \mathbf{I}_W)\boldsymbol{s}_r = \mathbf{P}_r\boldsymbol{s}_r$, where where $\boldsymbol{p}_r \in \mathbb{R}^{N_r}$ represents the set of panchromatic pixels of region $r$, $N_r$ being the number of pixels in region $r$; $\mathbf{I}_W \in \mathbb{R}^{W\times W}$ is an identity matrix; $\boxtimes$ denotes a Kronecker product; and $\boldsymbol{s}_r \in \mathbb{R}^W$ is the reference spectrum for this region.

The acquisition model for this region can be expressed as: $\boldsymbol{d}_r = \mathbf{H}_r\boldsymbol{o}_r + \boldsymbol{n}_r = \mathbf{H}_r\mathbf{P}_r\boldsymbol{s}_r + \boldsymbol{n}_r = \mathbf{G}_r\boldsymbol{s}_r + \boldsymbol{n}_r$. The estimation of the reference spectrum $\boldsymbol{s}_r$ is formulated as an optimization problem such as:

$$\widehat{\boldsymbol{s}}_r = \underset{\boldsymbol{s}_r>0}{\operatorname{argmin}} \|\boldsymbol{d}_r - \mathbf{G}_r\boldsymbol{s}_r\|^2_{\mathbf{\Gamma}^{-1}} + \mu\|\mathbf{D}\boldsymbol{s}_r\|^2, \qquad (3)$$

where $\mathbf{\Gamma} = \operatorname{diag}\{\boldsymbol{d}_r\}$ is the covariance matrix of the noise (Poisson noise approximated as a Gaussian noise), and $\mu\|\mathbf{D}\boldsymbol{s}_r\|^2$ is a Tikhonov regularization that quadratically penalizes the derivative along the spectral dimension. Such a solution can be computed easily using a quadratic programming solver.

However, using such a method to estimate a reference spectrum $\widehat{\boldsymbol{s}}_r$ requires to guarantee that the Separabilty Assumption is valid for region $r$.

### 2.2. Intraclass spectral variability model

In terms of spectral classification, the previous Separabilty Assumption amounts to consider the simple intra-class spectral variability model proposed by [4], as spectra of pixels in the same class only differ by a multiplicative coefficient.

With such a model, if pixel $n$ is of class $k$ with a reference spectrum $\boldsymbol{s}_k$, its spectrum $\boldsymbol{o}_n$ can be expressed as $\boldsymbol{o}_n = \psi_n\boldsymbol{s}_k$, where $\psi_n$ is the intra-class spectral variability coefficient for pixel $n$. Such a coefficient can be estimated easily from the panchromatic image, as the panchromatic value $p_n$ of pixel $n$ corresponds to the integration of its spectrum over all wavelengths: $W\,p_n = \mathbb{1}_W^T\boldsymbol{o}_n = \psi_n\mathbb{1}_W^T\boldsymbol{s}_k = \psi_n\widetilde{s}_k$, where $\mathbb{1}_W \in \mathbb{R}^W$ denotes a unit vector, and $\widetilde{s}_k = \sum_w \boldsymbol{s}_k(w)$. Thus, $\psi_n$ can be estimated simply by $\widehat{\psi}_n = \widetilde{s}_k^{-1}p_n$. Note that it is also possible to estimate $\widehat{\psi}_n$ from the coded data $\boldsymbol{d}_n$ using the relation $\boldsymbol{d}_n = \mathbf{H}_n\boldsymbol{o}_n + \boldsymbol{n}_n = \psi_n\mathbf{H}_n\boldsymbol{s}_k + \boldsymbol{n}_n$, however, as demonstrated in previous studies [5], estimation from the panchromatic image has yielded better results.

Such a model and estimator of $\widehat{\psi}_n$ can be used to test whether pixel $n$ is of class $k$ or not.

### 2.3. Gaussianity statistical tests

In the context of coded hyperspectral data, the instrument noise is often modeled as Poisson noise [6]. However, because the coded aperture system tends to reduce intensity variations across the scene, we can approximate this Poisson noise as white Gaussian noise for each pixel. This assumption holds particularly well when considering limited homogeneous regions of the scene. This allows us to use Gaussianity statistical tests to check the Separability Assumption for a given region and associated spectrum, as the residuals $\boldsymbol{d}_r - \mathbf{G}_r\widehat{\boldsymbol{s}}_r$ should be Gaussian, or to test whether a pixel $n$ is in a given class $k$ as the residuals $\boldsymbol{d}_n - \psi_n\mathbf{H}_n\boldsymbol{s}_k$ should be Gaussian.

A statistical test is used to verify a hypothesis about a random variable, such as its statistical parameters or its fit to a given probability distribution, with a certain level of uncertainty called the significance level $\alpha$. By default, $\alpha$ is set at

0.05, meaning that a 5% uncertainty is allowed in the test results, providing 95% confidence in the conclusions.

There exists a family of tests dedicated to verifying the Gaussianity of a random variable. Among these are parametric tests like the Kolmogorov-Smirnov (KS) test and non-parametric tests like the Shapiro-Wilk (SW) test [7]. Theoretically, the SW test is regarded as the most powerful among Gaussianity tests, surpassing options like KS, Anderson-Darling, and Lilliefors [8]. Moreover, an essential distinction is that the KS test is requiring a large sample size to yield significant results, while the SW test is valid for any random variable sizes as small as six elements, making it better suited to our data constraints. In practice, we have tested and compared these two tests, with the SW test consistently outperforming the KS test. For these reasons, we have selected the SW test for the remainder of our work.

### 2.4. Proposed algorithm

In study [5], we demonstrated the relevance of the separability hypothesis [3] and Gaussian white additive noise for real coded data. This approach was validated using statistical tests in the context of supervised classification, assuming the reference spectra of the scene materials were known. However, in practice, these reference spectra are not available.

Our recent studies have shown the effectiveness of using non-parametric Gaussianity statistical tests, such as the SW test, for detecting homogeneous regions from which we can extract reference spectra. This approach involves an SA procedure to estimate the spectrum of a selected region, followed by a prediction of the coded data in this region. The tests are then applied to the residuals between the real and predicted data.

By combining the SA method with statistical tests, we propose a new unsupervised classification approach that utilizes spatial characteristics calculated from the panchromatic image such as light intensity, alongside Gaussianity tests to validate the estimation of reference spectra and determine the class.

We propose a 3-steps approach, starting with a **preprocessing**, followed by a loop of **detection** and one of **labeling**. The detailed procedure is illustrated in Fig. 2.

The proposed unsupervised classification algorithm for coded data consists of three main steps: **preprocessing** based on the panchromatic image, **detection** of homogeneous regions and extraction of reference spectra, and **labeling** of all pixels belonging to the same material as the homogeneous detected region. The algorithm is designed to stop when all remaining unlabelled pixels are not validated as good candidates for a new homogeneous region. Additional stopping conditions include a maximum number of iterations to prevent excessive computations.

The **preprocessing** step aims to consider the spatial aspect of the data, particularly by utilizing the panchromatic

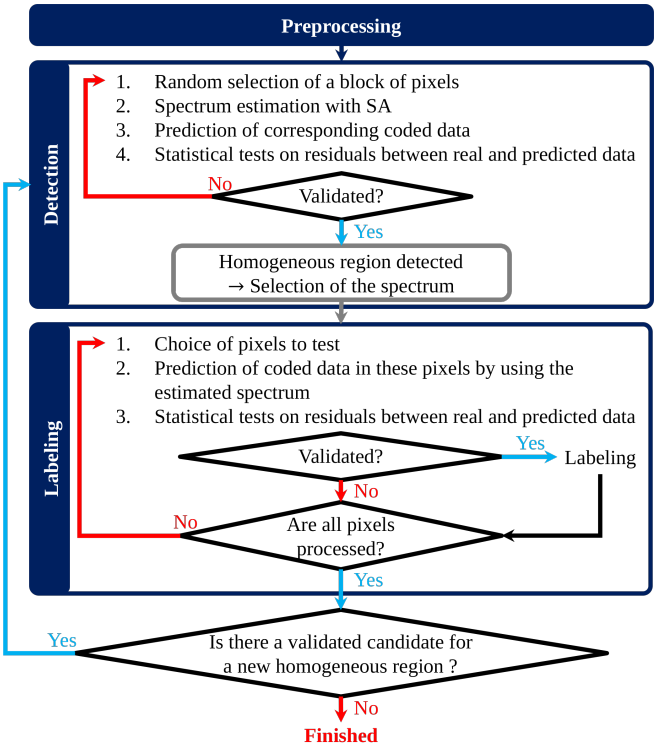


**Fig. 2**. Unsupervised classification of coded data algorithm

image. This involves thresholding the panchromatic image to remove low-intensity regions that do not contain sufficient information or with too low signal to noise ratio for accurate SA reconstruction. The threshold value is crucial and must be carefully chosen to eliminate dark pixels while retaining enough detail in the rest of the scene.

A homogeneous region is **detected** using statistical tests applied to the residuals between the real and predicted coded data. A square block of dimensions $\mathbb{R}^{\sqrt{P}\times\sqrt{P}}$ is defined around a central pixel randomly chosen among the preprocessing results. The SW test is used to verify the gaussianity of the residuals, assuming Gaussian white noise in the coded data like mentionned the previous section 2.1. If the test confirms gaussianity, the region is considered homogeneous and the estimated spectrum $\widehat{\boldsymbol{s}}_r$ is considered to be a reference spectrum $\boldsymbol{s}_k$ for the current class $k$.

The **labeling** step assigns labels to all pixels belonging to the same material as the detected homogeneous region.

## 3. SIMULATION AND DISCUSSIONS

### 3.1. Simulation context

The algorithm was applied to both simulated and real datasets. The *LEGO Wall* dataset (simulated dataset) was generated based on a scene featuring a wall of LEGO bricks from [9] (Fig. 3.a), while the *Countryside* dataset (real dataset) is an extraction from the CAMCATT dataset [10] (Fig. 3.b). De-

tailed characteristics of the datasets are provided in Table 1.

| | *LEGO Wall* | *Countryside* |
|---|---|---|
| **Dimensions** $(R, C, W)$ | $397, 399, 110$ | $451, 351, 117$ |
| **Wavelengths** [nm] | 400 - 700 | 380 - 780 |
| **SNR** [dB] | $\sim 30$ | Unknown |
| **Noise distribution** | *iid.* gaussian | Unknown |
| **PSF** *(Point Spread Function)* | 2D Gaussian $(\sigma = 2, \forall w)$ | Unknown |

**Table 1**. Dataset characteristics

On the simulated dataset, a Gaussian PSF with a standard deviation of $\sigma = 2$ was artificially applied to all wavelengths $w$ of the hyperspectral scene $\boldsymbol{o}$, while on the real dataset, the PSF is unknown. This implies that homogeneous regions are well-defined in the *LEGO Wall* dataset, whereas in the *Countryside* dataset, they are visually estimated (for example, each field could be considered a homogeneous region).

The corresponding coded data for both datasets were simulated using the SIMCA simulator [11] with DMD masks in a normalized orthogonal length-S configuration, which is considered the most optimized configuration [12]. An additive Gaussian white noise with known parameters was added on coded data from *The Wall* dataset.

In both cases, for the classification process, the signification level of the statistical test was set as a typical value of $\alpha = 0.05$, which means we tolerate a $5\%$ error rate.

The objective is to label as many homogeneous regions as possible. Non-classified regions could be considered as mixed regions, because of the PSF, or as undetermined classes due to the lack of information.

### 3.2. Classification results

Figures 3.c and 3.d depict the classification results obtained by applying our algorithm to the coded data computed from the *LEGO Wall* and *Countryside* datasets, respectively.

In these figures, black-colored labels represent pixels that were eliminated in the preprocessing step. Pixels that remain unclassified are labeled in white. Both black and white labels are considered as non-classified. Otherwise, each label is represented by a random color.

For the *LEGO Wall* dataset (Fig. 3.b), our algorithm successfully detected 19 classes automatically, without prior knowledge of the number of classes (ground truth: 21 classes after intensity threshold preprocessing, by considering each brick as a class for simplicity). It is evident that the regions inside the bricks were mostly correctly labeled, while those at the frontiers between bricks remained unclassified. This outcome is intended, as it is a feature of our algorithm to reject pixels located near the frontier between two homogeneous regions. Indeed, these pixels likely contain a mixture of spectra due to the PSF. However, some frontiers between bricks were not rejected (for example, in the red selected region between

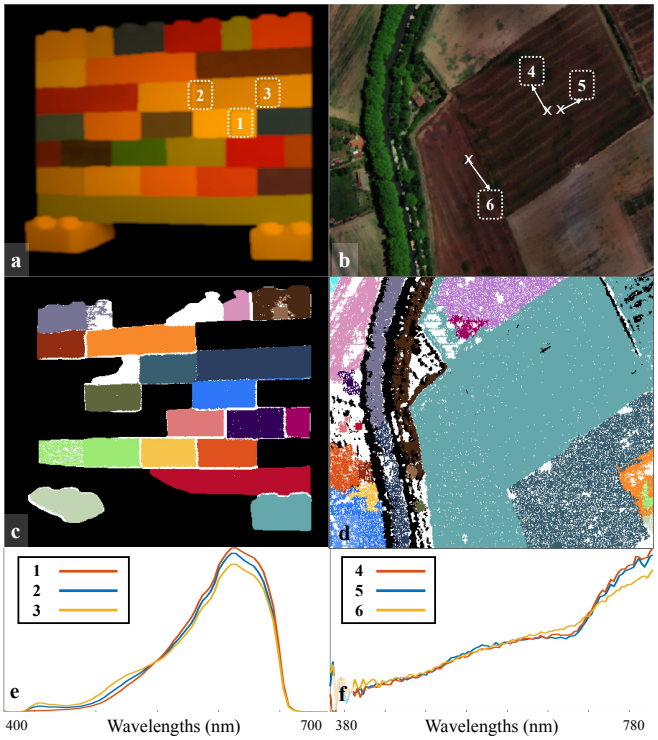


**Fig. 3**. Unsupervised classification results – **a**: False-color scene of *LEGO Wall* ; **b**: False-color scene of *Countryside* ; **c**: Classification results from *LEGO Wall* ; **d**: Classification results from *Countryside* ; **e**: Normalized spectra from regions 1, 2 and 3 from (a) ; **f**: Normalized spectra from pixels 4, 5 and 6 from (b)

bricks 1 and 2), which is understandable when considering that the reference spectra of these bricks are extremely similar (see Fig. 3.e).

In the case of the *Countryside* dataset, since the noise distribution is unknown, it is challenging to choose an appropriate statistical test. We decided to use the same SW test as in the case of the *LEGO Wall* dataset, assuming the noise to be white Gaussian. In Fig. 3.d, different large fields were mostly correctly identified by different labels. Some regions with complex details were ignored, such as regions combined with trees, houses, and grass. Other parts of the fields with different types of soil were also considered as the same class because the corresponding reference spectra are very similar (Fig. 3.f). However, there are some noticeable imperfections, for example, the green tree zone on the left of the scene was divided into multiple labels instead of one or two labels, and the grass zone and the field zone at the bottom-left corner of the scene were not correctly separated.

### 3.3. Discussions and perspectives

In this study, we presented a promising approach for the unsupervised classification of coded hyperspectral data without

reconstructing the cube. Despite having only a tenth of the information, an unknown number of materials, and the absence of reference spectra, we successfully detected most homogeneous regions of the observed scene while effectively rejecting mixed regions, both in simulated and real datasets.

However, there is still room for future improvements. We aim to exploit other spatial characteristics of the panchromatic image beyond intensity level to apply to all three steps of the algorithm. For the preprocessing step, spatial characteristics could be used to eliminate regions that are probably not homogeneous, for example, by presenting different textures than their neighboring regions. For the detection step, our preliminary studies showed that the classification results depend on the initialization order. By incorporating more spatial characteristics, we aim to establish a processing order that could be applied to all scenes uniformly. For the labeling step, we intend to add spatial constraints to precisely determine region boundaries.

For the coded data simulated from real dataset, further statistical studies will be necessary to determine a correct noise model, and appropriate statistical tests could be utilized accordingly. Lastly, a class fusion step could be integrated into our algorithm by comparing estimated reference spectra of existing classes or by employing other statistical tests (such as group variances tests) to verify the invariance of data values between different classes.

Additionally, further validation of statistical tests is necessary. This includes examining type II errors (false negative) and performing tests on residuals between real and predicted data to understand how confidence in the test decreases as the distance from the detected homogeneous region increases. This will provide a pseudo-probability map, offering better insights into the test's behavior relative to real data. Spatial aspects of the data must also be considered when estimating the variability coefficient, correlating the p-value, the estimated $\psi$ value, and the spatial distance.

Regarding unclassified regions, our long-term goal is to develop adaptive strategies to adjust the mask configurations based on the results of this homogeneous region detection phase to obtain complementary data for a future comprehensive scene analysis.

## 4. REFERENCES


[1] M. E. Gehm, R. John, and D. J. Brady, "Single-shot compressive spectral imaging with a dual-disperser architecture," *Optics Express*, vol. 15, no. 21, pp. 1013–14027, Nov. 2007.

[2] H. Zhang, X. Ma, X. Zhao, and G. R. Arce, "Compressive spectral image classification using 3D coded convolutional neural network," *Optics Express*, vol. 29, no. 21, pp. 32875–32891, 2021.

[3] I. Ardi, *Reconstruction d'images pour un imageur hyperspectrale configurable*, Ph.D. thesis, Université Toulouse III - Paul Sabatier, Toulouse, France, Oct. 2020.

[4] L. Drumetz, *Endmember Variability in hyperspectral image unmixing*, Signal and Image processing, Université Grenoble Alpes, Université Grenoble Alpes, Oct. 2016.

[5] T.-T. Dinh, H. Carfantan, A. Monmayrant, and S. Lacroix, "Tests statistiques pour l'analyse d'acquisitions hyperspectrales codées," in *XXIXème Colloque Francophone de Traitement du Signal et des Images*, Grenoble, 2023.

[6] F. Deger, A. Mansouri, M. Pedersen, J. Yngve Hardeberg, and Y. Voisin, "A Variational Approach for Denoising Hyperspectral Images Corrupted by Poisson Distributed Noise," in *ICISP 2014: Lecture Notes in Computer Science*. 2014, vol. 8509 of *International Conference on Image and Signal Processing*, pp. 106–114, Springer, Cham.

[7] S. S. Shapiro and M. B. Wilk, "An analysis of variance test for normality (complete samples)," *Biometrika*, vol. 52, no. 3-4, pp. 591–611, Dec. 1965.

[8] N. M. Razali and B. W. YAP, "Power Comparisons of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors and Anderson-Darling Tests," *Journal of Statistical Modeling and Analytics*, vol. 2, no. 1, pp. 21–33, 2011.

[9] E. Hemsley, I. Ardi, T. Rouvier, S. Lacroix, H. Carfantan, and A. Monmayrant, "Fast reconstruction of hyperspectral images from coded acquisitions using a separability assumption," *Journal of the Optical Society of America A*, vol. 30, no. 5, pp. 8174–8185, Feb. 2022.

[10] L. Roupioz, X. Briolett, K. Adeline, A. Al Bitar, D. Barbon-Dubosc, R. Barda-Chatain, P. Barillot, S. Bridier, E. Carroll, and C. Cassa, "Multi-source datasets acquired over Toulouse (France) in 2021 for urban microclimate studies during the CAMCATT/AI4GEO field campaign," *Data in Brief*, vol. 48, pp. 109, 2023.

[11] A. Rouxel, A. Monmayrant, S. Lacroix, H. Camon, and S. Lopez, "Accurate ray-tracing optical model for coded aperture spectral snapshot imagers," *Applied Optics*, vol. 63, no. 7, pp. 1828–1838, 2024.

[12] E. Hemsley, I. Ardi, S. Lacroix, H. Carfantan, and A. Monmayrant, "Optimized coded aperture for frugal hyperspectral image recovery using a dual-disperser system," *Journal of the Optical Society of America A*, vol. 37, no. 12, pp. 1916–1926, Nov. 2020.