# Toulouse Hyperspectral Data Set: a benchmark data set to assess semi-supervised spectral representation learning and pixel-wise classification techniques

Romain Thoreau*[1], Laurent Risser[2], Véronique Achard[3], Beatrice Berthelot[4], and Xavier Briottet[3]

[1]Centre National d´tudes Spatiales [Toulouse] – Centre National d'Etudes Spatiales - CNES (Toulouse, France) – France
[2]Institut de Mathématiques de Toulouse UMR5219 – Université Toulouse Capitole, Institut National des Sciences Appliquées - Toulouse, Université Toulouse - Jean Jaurès, Université Toulouse III - Paul Sabatier, Centre National de la Recherche Scientifique – France
[3]ONERA, Université de Toulouse [Toulouse] – PRES Université de Toulouse, ONERA – France
[4]Magellium – Magellium, Ramonville Saint-Agne, France – France

**Résumé**

Airborne hyperspectral images can be used to map the land cover in large urban areas, thanks to their very high spatial and spectral resolutions on a wide spectral domain. While the spectral dimension of hyperspectral images is highly informative of the chemical composition of the land surface, the use of state-of-the-art machine learning algorithms to map the land cover has been dramatically limited by the availability of training data. To cope with the scarcity of annotations, semi-supervised and self-supervised techniques have lately raised a lot of interest in the community. Yet, the publicly available hyperspectral data sets commonly used to benchmark machine learning models are not totally suited to evaluate their generalization performances due to one or several of the following properties: a limited geographical coverage (which does not reflect the spectral diversity in metropolitan areas), a small number of land cover classes and a lack of appropriate standard train / test splits for semi-supervised and self-supervised learning.
Therefore, we release in this paper the Toulouse Hyperspectral Data Set that stands out from other data sets in the above-mentioned respects in order to meet key issues in spectral representation learning and classification over large-scale hyperspectral images with very few labeled pixels. The Toulouse Hyperspectral Data Set is the combination of 1) an airborne hyperspectral image acquired by the AisaFENIX sensor over Toulouse, France, during the CAMCATT-AI4GEO campaign (Roupioz et al., 2023) and of 2) a land cover ground truth, provided with standard train / test splits to foster fair and reproducible benchmarks. The image is provided in ground-level reflectance with a very high spatial resolution (1m ground sampling distance) and spectral resolution ($<$ 8 nm) from 0.4 $\mu$m to 2.5 $\mu$m (310 channels). More than 380,000 pixels are sparsely labeled over an area of 90 km2 . The land cover nomenclature contains 32 classes hierarchically organized into 16 impermeable surfaces and 16 permeable surfaces. As much as intra-class spectral shifts are correlated to the geographical location of pixels, we fostered the statistical independence of the training, validation and test sets by separating them geographically.
We discuss the differences between the Toulouse data set and other public hyperspectral

*Intervenant

data sets, showing its interest for the evaluation of machine learning models generalizability. In particular, we perform a qualitative comparison of the Toulouse data set with Pavia University and Houston University data sets, based on a hand-crafted representation technique. The qualitative comparison illustrates the larger spectral variability of the Toulouse data set, that is more representative of large urban areas. Finally, we discuss and experiment several self-supervised learning techniques and establish a baseline for pixel-wise classification. The Toulouse Hyperspectral Data Set is publicly available at www.toulouse-hyperspectral-data-set.com.