End-to-End Learning for Land Cover Classification Using Irregular and Unaligned SITS by Combining Attention-Based Interpolation With Sparse Variational Gaussian Processes

Valentine Bellet 🆔, *Student Member, IEEE*, Mathieu Fauvel 🆔, *Senior Member, IEEE*, Jordi Inglada 🆔, and Julien Michel 🆔

Abstract—In this article, we propose a method exploiting irregular and unaligned Sentinel-2 satellite image time series (SITS) for large-scale land cover pixel-based classification. We perform end-to-end learning by combining a time and space informed kernel interpolator with a sparse variational Gaussian processes (SVGP) classifier. The interpolator embeds irregular and unaligned SITS onto a fixed and reduced size latent representation. The spatial information is taken into account by using a *spatial positional encoding*. The obtained latent representation is given to the SVGP classifier and all the parameters are jointly optimized w.r.t. the classification task. We run experiments with irregular and unaligned Sentinel-2 SITS of the full year 2018 over an area of 200 000 km^2 (about two billion pixels) in the south of France (27 MGRS tiles). Such experimental condition exacerbates the irregular and unaligned issues of SITS. In terms of overall accuracy, with the learned latent representation instead of linearly interpolated SITS, the results of the SVGP classifier are improved by about 10 points. Moreover, with the learned latent representation, the SVGP classifier outperforms the main state-of-the-art methods from the literature at large scale (e.g., seven points for the multilayer perceptron) and is robust to the available timestamps used for training and testing.

Index Terms—Land cover map, large-scale classification, representation learning, satellite image time-series (SITS), Sentinel-2, sparse variational Gaussian processes (SVGP).

I. INTRODUCTION

IN MARCH 2023, the final synthesis report of the sixth assessment report was released by the intergovernmental panel on climate change. Its main conclusions are that climate impacts on ecosystems are more intense and widespread than expected [1]. Among other recommendations, they proposed to expand the use of digital technology for land use monitoring and sustainable land management, which can help to reduce emissions from deforestation and land-use changes.

Earth observation (EO) satellites provide a huge amount of raw data of different types (e.g., optical or radar). Extracting meaningful information from these raw EO data enables the monitoring of the Earth's surface changes and therefore can help to solve the challenges of climate change [2], [3]. For instance, the Sentinel-2 twin satellites provide free and open-access data with relevant features: short revisit time (five days) and high spectral and spatial resolutions (four spectral bands at 10 m, six at 20 m, and three at 60 m per pixel) [4].

These satellite image time-series (SITS), covering large continental surfaces with a short revisit cycle, bring the opportunity of large-scale mapping. For example, land use or land cover (LULC) maps provide information about the physical and functional characteristics of the Earth's surface for a particular period of time. More precisely, land cover usually refers to the physical land type (i.e., corn field or grassland), whereas land use map indicates how people are using the land (i.e., agriculture). To produce these LULC maps from massive SITS, automatic methods are mandatory. In the last years, machine learning (ML) and then deep learning (DL) methods have shown outstanding results in terms of performance accuracy [5], [6], [7].

A widely used ML algorithm for pixel-wise classification, with very good performances even in large scale, is the random forest (RF) [8], [9], [10]. However, this classifier is not able to take into account the spectro-temporal structure of the SITS. In recent years, DL methods have been developed and have shown very accurate results. Indeed, they are able to extract features (i.e., spatial, spectral, or/and temporal) of the SITS. For example, a combination of convolutional neural network (CNN) and recurrent neural network (RNN) has shown good performances by including the spatial information [11]. Temporal CNN, by combining temporal and spatial features, have also shown satisfactory results [12]. Recently, methods based on attention mechanisms were proposed in order to take into account the spectro-temporal structure of the data [13]. However, DL methods have a huge number of parameters, which are sometimes difficult to interpret and to optimize. Recently, we proposed a method based on sparse variational gaussian processes (SVGP). This method takes into

account the spatio-spectro-temporal structure of the data through a covariance function and its parameters are interpretable. It provides similar classification performance to the state-of-the-art methods, such as conventional ML or DL methods. However, we identified two limiting factors: the need of preimputation or resampling of the data in the temporal domain, like most of the methods in literature which require data with a constant number of features per pixel, and the high number of spectro-temporal features that can perturb the optimization process.

Unfortunately, Sentinel-2 pixel time series are irregularly sampled in the temporal domain: observations are not equally spaced in time due to the presence of clouds or shadows. These time series are also unaligned: observations from two different satellite swaths have different temporal sampling grids. Preprocessing techniques can be used to transform these irregular and unaligned time series into regular time series that can be used by the classifier. For instance, Inglada et al. [9] proposed to linearly resample the observations onto a common set of latent dates. The obtained resampled observations from a full year were successfully used to produce land cover classification maps at country scale using SVGP [14]. However, relevant information for the classification task can be lost when producing these resampled observations. Indeed, Li et al. [15] showed that an independent interpolation method directly followed by a classification method performed worse than methods trained end-to-end.

In this sense, Constantin et al. [16] proposed to jointly classify and reconstruct irregular pixel time series. Despite the quality of the reconstruction, the model did not compete with state-of-the-art classifiers, such as RF or support vector machine (SVM) because of too strong statistical assumption. Besides, Petitjean et al. [17] proposed to use dynamic time warping (DTW). DTW allows to find the best alignment between two time series, however, it does not include information on inter- and intraannual phenological cycles. [18]. Thus, the time-weighted dynamic time warping (TWDTW) was proposed by introducing time weight factor, as an extension of the DTW [19]. Later, a parallel version of the TWDTW was proposed, taking into account the spatial dimension [20]. Even if it achieved almost linear speed up, it was not able to deal with very large datasets.

Few DL methods can directly deal with these irregular and unaligned time series. For example, long short-term memory (LTSM) [21] can take into account irregular time series, however, they do not support unaligned time series. In land cover classification, Ienco et al. [22] used LSTM combined with linear interpolation in order to deal with missing (i.e., cloudy) observations. Moreover, LSTM are slow to train because of the lack of parallelization abilities. Transformer architectures [23], via the self-attention mechanism, are able to process sequences in parallel, and dealing with irregular and unaligned time series is done via temporal positional encoding and padding. Rußwurm and Körner [24] pioneered the use of self-attention for land cover mapping using Sentinel-2 SITS. Garnot et al. [13] improved the approach by reducing the computational complexity with the lightweight temporal self-attention (LTAE). The method outperforms most of state-of-the-art time series classification algorithms. However, these DL methods still require a huge number of parameters, which are often not interpretable.

To take advantage of the abovementioned SVGP approach [14], we propose to learn a fixed-size latent representation as a preprocessing step to the classifier. This strategy was explored in [25] and [26], where the authors proposed a method called multi-time attention networks (mTAN), which enables working with irregular and unaligned time series. By using end-to-end training, a learned kernel interpolator (the mTAN) followed by an encoder–decoder task provided accuracies similar to or better than the state-of-the-art for a classification task [26]. Although kernel interpolators have been known for a long time, the improved learning capacity of mTAN results from the temporal attention used as the similarity kernel learned from the data.

In this work, we propose to further elaborate on a learned kernel interpolator for the classification of irregular and unaligned SITS. Our first contribution is to propose a time and space informed kernel interpolator based on the mTAN for solving the issues identified in our previous work with SVGP. This module learns a constant-size latent representation from irregular and unaligned multivariate time series, in a end-to-end learning framework with SVGP. The mTAN is modified to

1) take into account the geographic coordinates of the pixel thanks to a spatial positional encoding,
2) perform a learned spectro–temporal feature reduction.

We also propose a formal interpretation of the resulting kernel operator: to the best of our knowledge, this is the first time that such a kernel operator is proposed for the classification of SITS.

Our second contribution is the evaluation of the proposed model with existing state-of-the-art methods from the literature (limited to those that can be used at large scale), on 27 Sentinel-2 tiles in the south of France. Such experimental condition exacerbates the irregular and unaligned issues of SITS, which have scarcely been analyzed in the geosciences and remote sensing community.

Beyond improved classification accuracy and reduced training time, the proposed model is versatile w.r.t. the temporal sampling: during inference it can classify any irregular and unaligned pixel time series even if its timestamps were not seen during the training process.

The rest of this article is organized as follows. Section II-A describes how the learned kernel interpolator (the mTAN) is used to process irregular and unaligned pixel time series. Section II-B defines our contributions for large scale land cover classification with irregular and unaligned SITS. The experimental setup is detailed in Section III. The results obtained with the end-to-end trained model (time and space informed kernel interpolator coupled with SVGP) are provided in Section IV. Comparison with existing state-of-the-art methods from the literature are provided in Section V. Latent representation and the similarity kernel learned by the interpolator are discussed in Section VI. Finally, Section VII concludes this article.

II. METHODS

This section describes how irregular and unaligned pixel time series are projected onto a fixed temporal grid in order to be used
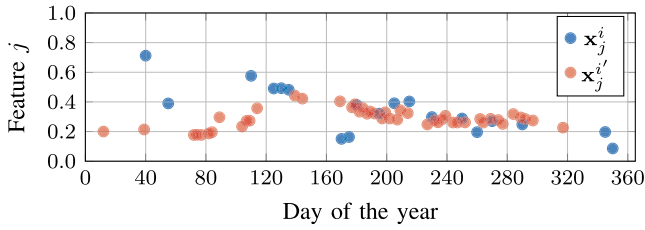
Fig. 1. $\mathbf{x}_j^i$ and $\mathbf{x}_j^{i'}$ are two irregular and unaligned time series for pixels i and i', respectively, for the spectral feature j.

by the classifier. First, some notations and definitions, which are used throughout this article are introduced. Then, the latent interpolator at the core of the proposed method is presented. Finally, the last part describes our several contributions.

A. Attention-Based Temporal Interpolator

1) Notations and Definitions: In this article, the ith pixel time series $\mathbf{x}^i(t_k)$ at time t_k is defined by its spectral measurements $\{x_1^i(t_k), \ldots, x_j^i(t_k), \ldots, x_D^i(t_k)\}$ with $i \in \{1, \ldots, N\}$, N the number of pixels and D the number of spectral features. In addition, two spatial coordinates ψ_1^i and ψ_2^i are associated to the pixel $\mathbf{x}^i$. Moreover, $y^i \in \{1, \ldots, C\}$ is the target value (i.e., the class label) associated to the pixel $\mathbf{x}^i$, with C the number of classes.

For a pixel i, a spectral feature j is observed at T_j^i timestamps: $\mathbf{T}_j^i = \{t_{j1}^i, \ldots, t_{jk}^i, \ldots, t_{jT_j^i}^i\}$, where T_j^i is the number of valid observations (e.g., no clouds or shadows). As discussed in Section I, because of satellite swaths and weather we usually have unaligned time series, i.e., $\mathbf{T}_j^i \neq \mathbf{T}_j^{i'}$. In this work, we assume that all spectral features are available for each timestamp, i.e., $\mathbf{T}_j^i = \mathbf{T}_{j'}^i = \mathbf{T}^i$. This is commonly the case when working with only one sensor, but the proposed method can be extended to multisource data straightforwardly. As an illustration, Fig. 1 represents two real irregular and unaligned pixel time series acquired by Sentinel-2. We define the set of all timestamps $\mathbf{T}$ such as

$$\mathbf{T} = \bigcup_{i=1}^N \mathbf{T}^i$$
$$= \{t_1, \ldots, t_k, \ldots, t_T\}$$

with T the total number of observations. For each pixel, we define a mask time series $\mathbf{m}^i \in \{0, 1\}^T$ such as

$$m^i(t_k) = \begin{cases} 1 & \text{if } t_k \in \mathbf{T}^i \\ 0 & \text{otherwise} \end{cases} \quad \forall t_k \in \mathbf{T} \tag{1}$$

which indicates whether the feature j of pixel i at time t_k is observed or not. We further define an *augmented* pixel time series $\mathbf{x}_j^{i*}$ as the pixel

$$x_j^{i*}(t_k) = \begin{cases} x_j^i(t_k) & \text{if } m^i(t_k) = 1 \\ 0 & \text{otherwise} \end{cases} \quad \forall t_k \in \mathbf{T}. \tag{2}$$

Using (1) and (2) will simplify the presentation of the interpolator in the following section.

2) Projection Onto a Regular-Temporal Grid: As previously described, most of the classifiers are not able to deal with irregular and unaligned time series. Thus, the core idea is to learn a mapping of these irregular and unaligned time series onto a regular temporal grid of R latent dates: $\mathbf{R} = \{r_1, \ldots, r_l, \ldots, r_R\}$. In this work, we focus on the well-established Nadaraya–Watson kernel smoother [27, Ch. 6], because it leads to an efficient interpolation as discussed in the next section.

For a given pixel time series $\mathbf{x}_j^*$, the interpolated $\hat{x}_j$ at latent timestamp r_l using a kernel smoother is given by:[1]

$$\hat{x}_j(r_l) = \frac{\sum_{t_k=t_1}^{t_T} K(r_l, t_k) m(t_k) x_j^*(t_k)}{\sum_{t_k'=t_1}^{t_T} K(r_l, t_k') m(t_k')} \tag{3}$$

with K some similarity kernel [27, Ch. 6]. Usually, the radial basis function (RBF) kernel is used $K(r_l, t_k) = \exp(d(r_l, t_k))$ with $d(r_l, t_k) = -\sigma^{-2}(r_l - t_k)^2$. From (3), $\hat{x}_j(r_l)$ is a convex combination of original pixel values, whose weights are computed using the kernel applied on the temporal domain. With a RBF kernel, the similarity is a decreasing function of the distance between two timestamps, whatever their location in the year. The parameter σ, learned from the training data, weights the temporal distance.

The performances of such method are strongly limited by the hand-crafted similarity kernel. A powerful extension is obtained using *attention* and *embedding* mechanisms, which are able to build more complex (anisotropic) kernels [28, Ch. 11]. In the following, the mTAN [26] is discussed as an extension of the kernel smoother to build the kernel interpolator for the classification model in our end-to-end training.

3) mTAN: To build the kernel interpolator, Shukla et al. [26] proposed using attention mechanisms and more precisely the scaled-dot product attention defined in [23]. First, a learnable time embedding function (named *temporal positional encoding*) ϕ is defined. It maps a given timestamp t onto a higher dimensional space of size E such as

$$\phi : \mathbb{R} \to \mathbb{R}^E$$

$$t \mapsto \phi(t) = \begin{bmatrix} \omega_1 t + \alpha_1 \\ \sin(\omega_2 t + \alpha_2) \\ \vdots \\ \sin(\omega_E t + \alpha_E) \end{bmatrix} \tag{4}$$

with ω_p and α_p, $p \in \{1, \ldots, E\}$, the learnable parameters.

Then, this embedding ϕ is used to construct the similarity kernel K used in (3) such as

$$d(r_l, t_k) = \frac{\phi(r_l)^\top \mathbf{W}_q^\top \mathbf{W}_k \phi(t_k)}{\sqrt{E}}$$

with $\mathbf{W}_q$ and $\mathbf{W}_k$ two learnable matrices of size $E \times E$, the indices q and k refer to *query* and *key* terms in the attention mechanism framework [23].

---

[1]For clarity, we consider only one pixel and we drop the index i in the remaining of this article.
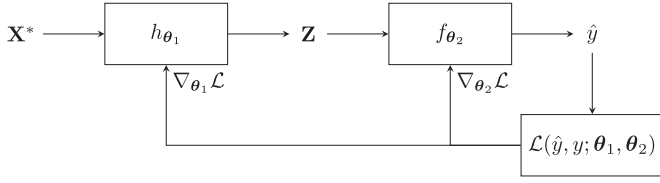
Fig. 2. End-to-end learning for the classification of one irregular and unaligned pixel time series $\mathbf{X}^*$ and its associated representation $\mathbf{Z}$.

Finally, (3) can be rewritten using a masked softmax operator [28, Ch. 11.3.2] such as

$$\hat{x}_j(r_l) = \text{softmax}\left\{\frac{\left(\Phi(\mathbf{T})^\top \mathbf{W}_k^\top \mathbf{W}_q \phi(r_l)\right) \odot}{\sqrt{E}}\right\}^\top \mathbf{x}_j^*$$
$$= \boldsymbol{\gamma}_{r_l}^\top \mathbf{x}_j^*. \tag{5}$$

with $\Phi(\mathbf{T}) = [\phi(t_1), \dots, \phi(t_T)]$, the matrix of embeddings of $\mathbf{T}$ and $\odot$ being the Hadamard product. $\mathbf{x}_j^*$ refers to *value* term attention mechanism framework [23].

Shukla and Marlin [26] further propose to use multihead attention, i.e., H matrices of embeddings with $\Phi_H(\mathbf{T}) = \{\Phi_h(\mathbf{T})\}_{h=1}^H$, and also H time embedding functions with $\phi_H(r_l) = [\phi_1(r_l), \dots, \phi_H(r_l)]$. A learnable linear layer $\boldsymbol{\beta}_H$ of size H is used to produce the interpolated value

$$\hat{x}_j(r_l) = \boldsymbol{\beta}_H^\top \boldsymbol{\Gamma}_{r_l}^{H^\top} \mathbf{x}_j^*. \tag{6}$$

with $\boldsymbol{\Gamma}_{r_l}^H = [\boldsymbol{\gamma}_{r_l}^1, \dots, \boldsymbol{\gamma}_{r_l}^H]$. This equation can be computed for every spectral feature j and every latent date r_l.

The mTAN, as defined in (6), has extended interpolation flexibility w.r.t. the conventional kernel smoother. Also, it is worth noting that (6) benefits from the computational efficiency of attention mechanism (parallel computation) and all parameters are learnable during the training step.

In [26], the mTAN was used as input and output layers in a encoder–decoder architecture and a classifier was jointly learned using feature from the latent-space. The next section describes how we extend the mTAN to use the spatial information and to reduce the spectral dimension of the SITS.

B. Spatially Informed Interpolator for Classification

Fig. 2 represents the workflow for the classification of one irregular and unaligned pixel time series $\mathbf{X}^*$ through its latent representation $\mathbf{Z}$. In this article, we propose to use end-to-end learning by combining a spatially informed interpolator $h_{\boldsymbol{\theta}_1}$ with the SVGP classifier $f_{\boldsymbol{\theta}_2}$ defined in [14]. The SVGP classifier uses kernel functions, i.e., RBF covariance functions, and no changes were made from [14] (i.e., same loss). Indeed, the loss $\mathcal{L}$ is used to optimize $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ (i.e., the parameters of $h_{\boldsymbol{\theta}_1}$ and $f_{\boldsymbol{\theta}_2}$, respectively) and to minimize the error between the predicted class $\hat{y}$ and the true class y.

1) Spectro-Temporal Feature Reduction: The mTAN interpolation allows to perform feature reduction, in the temporal domain, in the spectral domain or in both of them. Indeed, the interpolated feature j is of size R and by taking $R < T$

we can perform a temporal feature reduction. We propose to add a linear layer after the interpolation in order to perform spectral feature reduction. Noting $\hat{\mathbf{x}}(r_l) \in \mathbb{R}^D$ the vector of all interpolated spectral features at timestamp r_l, $\mathbf{B}$ a matrix of size $D' \times D$ with $D' \leq D$, the final latent interpolated pixel $\mathbf{z}(r_l)$ can be written as

$$\mathbf{z}(r_l) = \mathbf{B}\hat{\mathbf{x}}(r_l) \tag{7}$$

The overall spectro–temporal feature reduction can be written as

$$\mathbf{Z} = \mathbf{B}\mathbf{X}^*\boldsymbol{\Gamma} \tag{8}$$

where $\mathbf{Z} = [\mathbf{z}(r_1), \dots, \mathbf{z}(r_R)] \in \mathbb{R}^{D' \times R}$, $\mathbf{X}^* = [\mathbf{x}^*(t_1), \dots, \mathbf{x}^*(t_T)] \in \mathbb{R}^{D \times T}$, and $\boldsymbol{\Gamma} = [\boldsymbol{\gamma}_{r_1}, \dots, \boldsymbol{\gamma}_{r_R}] \in \mathbb{R}^{T \times R}$.

As defined in (8), the matrix $\boldsymbol{\Gamma}$ does not depend on the spectral features and the matrix $\mathbf{B}$ does not depend on time. Thus, as in Constantin et al. [16], the temporal reconstruction does not depend on the spectral features and the spectral feature reduction does not depend on the time. This constrained spectro–temporal structure reduces the complexity (number of parameters) of the model.

Yet, the spatial information is not taken into account. In the following section, we discuss how the spatial coordinates are integrated in the processing by means of spatial positional encoding.

2) Spatial Positional Encoding: We propose to add the spatial information in the estimation of $\mathbf{Z}$ by using a *spatial positional encoding*. As in [29], the spatial coordinates (ψ_1, ψ_2) are mapped onto a higher dimensional space of dimension F using φ

$$\varphi : \mathbb{R}^2 \to \mathbb{R}^F$$
$$(\psi_1, \psi_2) \mapsto \varphi(\psi_1, \psi_2)$$
$$= \left[\sin(\psi_1\nu_1), \cos(\psi_1\nu_1), \dots, \cos(\psi_2\nu_{F/4})\right]^\top$$

with $\nu_q = 10000^{-(2l)/F}$ and $q \in \{1, \dots, F/4\}$. $\varphi(\psi_1, \psi_2)$ is then given to a two-layer perceptron with ReLu activation functions to obtain a vector of size D, which is finally duplicated for each timestamp to get a spatial positional encoding matrix $\mathbf{P}$ of the same shape as $\mathbf{X}^*$ (i.e., $D \times T$). This matrix is added to the raw input data $\mathbf{X}^*$ before the spectro–temporal interpolation

$$\tilde{\mathbf{X}}^* = \mathbf{X}^* + \mathbf{P}. \tag{9}$$

The parameters of the perceptron are jointly optimized with the time and space informed kernel interpolator and the SVGP during the learning step.

The SVGP classifier $f_{\boldsymbol{\theta}_2}$ uses a kernel function over the latent spectro–temporal representations of two pixels, respectively, noted $\mathbf{Z}^i$ and $\mathbf{Z}^{i'}$ defined as

$$k(\mathbf{Z}^i, \mathbf{Z}^{i'}) = \exp\left(-\frac{\|\mathbf{Z}^i - \mathbf{Z}^{i'}\|_F^2}{2\ell^2}\right)$$

with $\|\cdot\|_F$ and $\langle\cdot,\cdot\rangle_F$ the Frobenius norm and inner product over matrices and ℓ the lengthscale parameter of the kernel. The

TABLE I
DESCRIPTION OF THE TIME AND SPACE INFORMED KERNEL INTERPOLATOR
PARAMETERS $\boldsymbol{\theta}_1$ AND THEIR CORRESPONDING SIZES

Parameters	Size
$\{\omega_p, \alpha_p\}_{p=1}^{E}$	$2(HE)$
$\mathbf{W}_q, \mathbf{W}_k$	$2(HE^2)$
$\mathbf{B}$	$D'D$
$\boldsymbol{\beta}_H$	H
MLP	$L_2(L_1 + D)$

The MLP corresponds to the parameters of a two-layer perceptron used to obtain the spatial positional encoding matrix P described in the previous section.

square Frobenius norm can be written as

$$\|\mathbf{Z}^i - \mathbf{Z}^{i'}\|_F^2 = \underbrace{\|\mathbf{B}\mathbf{X}^{i*}\boldsymbol{\Gamma}^i - \mathbf{B}\mathbf{X}^{i'*}\boldsymbol{\Gamma}^{i'}\|_F^2}_{A}$$

$$+ \underbrace{\|\mathbf{B}\mathbf{P}^i\boldsymbol{\Gamma}^i - \mathbf{B}\mathbf{P}^{i'}\boldsymbol{\Gamma}^{i'}\|_F^2}_{B}$$

$$+ 2\underbrace{\langle\mathbf{B}(\mathbf{X}^{i*}\boldsymbol{\Gamma}^i - \mathbf{X}^{i'*}\boldsymbol{\Gamma}^{i'}), \mathbf{B}(\mathbf{P}^i\boldsymbol{\Gamma}^i - \mathbf{P}^{i'}\boldsymbol{\Gamma}^{i'})\rangle_F}_{C}.$$

Terms A and B correspond to the distance between two pixels for spectro-temporal latent variables and for spatial latent variables, respectively. Term C corresponds to an interaction term between spectro-temporal and spatial latent variables. By comparison to our previous works [14], the covariance function $k(\mathbf{Z}^i, \mathbf{Z}^{i'})$ is composed of an additional element: the interaction term C, and in addition, the spatial distance is learned. With formulation (9), we have a supplementary source of information that links spectro-temporal and spatial terms. In the following section, the complexity of the model is discussed.

3) Model Complexity: The parameters $\boldsymbol{\theta}_1$ of the time and space informed kernel interpolator $h_{\boldsymbol{\theta}_1}$ and their corresponding sizes are summarized in Table I and the total number of learnable parameters is given by

$$2HE(1 + E) + DD' + H + L_1(L_2 + D).$$

As a reminder, the parameters $\boldsymbol{\theta}_2$ of the SVGP classifier $f_{\boldsymbol{\theta}_2}$ were highly dependent on the number of spectro-temporal features $T \times D$. By using an end-to-end training with the time and space informed kernel interpolator, this number is significantly reduced to $R \times D'$ with $R < T$ and $D' < D$ and therefore the total number of parameters $\boldsymbol{\theta}_2$ is reduced as well. Numbers of parameters for our dataset are given in the experimental Section IV-B.

III. DATASET AND EXPERIMENTAL SET-UP

The study area covers approximately $200\,000\,\text{km}^2$ in the south of metropolitan France. It is composed of 27 Sentinel-2 tiles, as displayed in Fig. 3.
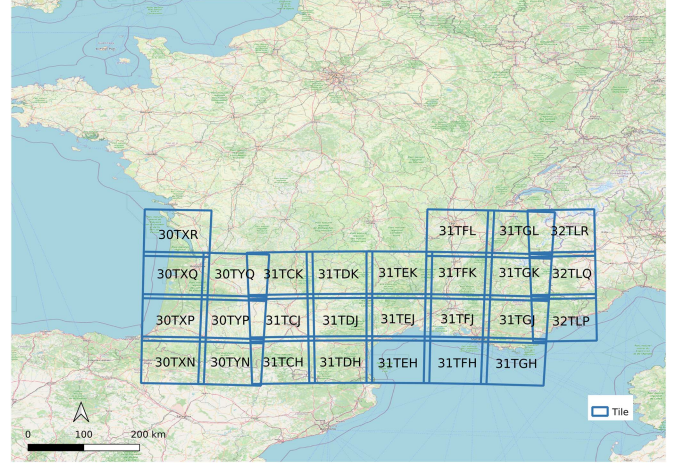

Fig. 3. Location of the 27 studied tiles where a blue square corresponds to one tile as provided by the Theia Data Center[2]. Each tile is displayed with its name in the MGRS nomenclature used for Sentinel-2 products (background map© OpenStreetMap contributors).

TABLE II
NUMBER OF PIXELS FOR EACH DATASET

Training	Validation	Test
92 000	23 000	230 000

A. Irregular and Unaligned Sentinel-2 Time Series

All available acquisitions of level 2 A between January and December 2018 for the 27 Sentinel-2 tiles were used, as described in [14]. Surface reflectance time-series and cloud/shadow masks have been produced using the MAJA preprocessing chain [30] and were downloaded from the Theia Data Center.[2] All the bands at 20 m/pixel were spatially up-sampled to 10 m/pixel using bicubic interpolation [31]. A total of 10 spectral bands with three spectral indices [normalized difference vegetation index (NDVI), normalized difference water index (NDWI) and Brightness] were used. Compared to [14], no temporal sampling preprocessing has been used (i.e., no linear interpolation as in [9] or other types of temporal synthesis). As described in Section I, the resulting data is irregular and unaligned. Following the notations defined in Section II-A3, the union of the acquisition dates of the 27 tiles results in $T = 303$ dates. Besides, the spectral dimension is equal to $D = 13$.

The reference data used in this work is composed of $C = 23$ land cover classes ranging from artificial areas to vegetation and water bodies constructed with different data sources as described in [14]. The nomenclature of the 23 land cover classes can be found in Table III.

Pixels were randomly sampled from polygons over the full study area (i.e., 27 tiles) to create three *spatially disjoint* data subsets: *training*, *validation,* and *test*. The polygons are disjoint between the three datasets. The three datasets are class-balanced:

[2][Online]. Available: https://www.theia-land.fr/en/products/

TABLE III
LAND COVER CLASSES USED FOR THE EXPERIMENTS WITH THEIR CORRESPONDING COLOR CODE

Color	Code	Name
	CUF	Continuous urban fabric
	DUF	Discontinuous urban fabric
	ICU	Industrial and commercial units
	RSF	Road surfaces
	RAP	Rapeseed
	STC	Straw cereals
	PRO	Protein crops
	SOY	Soy
	SUN	Sunflower
	COR	Corn
	RIC	Rice
	TUB	Tubers / roots
	GRA	Grasslands
	ORC	Orchards and fruit growing
	VIN	Vineyards
	BLF	Broad-leaved forest
	COF	Coniferous forest
	NGL	Natural grasslands
	WOM	Woody moorlands
	NMS	Natural mineral surfaces
	BDS	Beaches, dunes and sand plains
	GPS	Glaciers and perpetual snows
	WAT	Water bodies

TABLE IV
AVERAGED OVERALL ACCURACIES (OA) FOR THE *MTANE-SVGP* AND *GAPFILLED-SVGP* MODELS (MEAN % ± STANDARD DEVIATION COMPUTED WITH NINE RUNS)

	mTANe-SVGP	Gapfilled-SVGP
OA	77.44 ± 0.15	67.25 ± 0.37

4000 pixels per class in the *training* dataset, 1000 pixels per class in the *validation* dataset, and 10 000 pixels per class in the *test* dataset. The total number of pixels for each dataset is provided in Table II. Classification metrics such as overall accuracy (OA) or F-score were computed for each model using the *test* dataset with nine runs with different random pixel samplings. Standardization was performed for the valid acquisitions dates. Mean and standard deviation were estimated for each spectral band and for each spectral index on the *training* dataset and then used to standardize the other datasets (*validation, test*) [32].

B. Competitive Methods

Linearly interpolated data was feed into a simple SVGP classifier called *Gapfilled-SVGP* model and is used as baseline to compare with the extended mTAN combined with the SVGP, called *mTANe-SVGP* model. Two other classifiers are also compared in terms of classification accuracy and processing time:
1) Multi-layer perceptron (MLP) with the same setup as in [14].
2) Lightweight temporal self-attention (LTAE) described in [13].

The end-to-end trained models are called *mTANe-SVGP*, *mTANe-MLP*, and *mTANe-LTAE*, respectively.

Unlike SVGP or MLP classifiers, the LTAE classifier uses attention mechanisms. It may be redundant to use attention mechanisms both in the time and space informed kernel interpolator and in the LTAE. Therefore, the LTAE classifier was also studied without the extended mTAN (mTANe) and this method is called *raw-LTAE*. However, the LTAE classifier was not defined to deal with the irregular and unaligned time series pixels. Thus, the mask was used as an additional feature. Besides, the spatial

positional encoding matrix $\mathbf{P}$ was also used in this classifier, as defined in (9).

The optimizer parameters (i.e., number of epochs, learning rate, and batch size) for each model were found by trial and error and are described in Table VIII in Appendix A. To train all models, one NVIDIA Tesla V100 GPU was used.

IV. MODEL EVALUATION

This section presents the different results obtained by the *mTANe-SVGP* model. First, a comparison with linear interpolation is provided. Then, the influence on the classification accuracy and processing time of latent representation sizes as well as the use of the spatial positional encoding matrix are investigated. Finally, the influence of the number of inducing points is studied.

A. Comparison With Linear Interpolation

First, the *mTANe-SVGP* was implemented with a vector of latent dates $\mathbf{R}$ defined with a regular sampling of $\tau = 10$ days and a total number of $R = 37$ dates.[3] The number of latent spectral features was equal to the number of spectral features such as $D' = D = 13$. The latent representation $\mathbf{Z}$ obtained using the extended mTAN is described by $R \times D' = 481$ spectro–temporal features. The *Gapfilled-SVGP* model was implemented with the same number of spectro–temporal features. A detailed evaluation of the *Gapfilled-SVGP* model was done in [14].

The comparison in terms of OA between the *Gapfilled-SVGP* and *mTANe-SVGP* models is given in Table IV. The F-score, recall, and precision per class for both models are represented in Appendix C. As shown in Table IV, the *mTANe-SVGP* model is 10 points above the *Gapfilled-SVGP* model in terms of classification accuracy. The learned latent representation $\mathbf{Z}$ obtained by the time and space informed kernel interpolator contains more meaningful information for the classification task for the SVGP classifier compared to the linearly interpolated data.

B. Spectral and Temporal Feature Reduction

Figs. 4 and 5 represent, respectively, the averaged OAs and the averaged training times computed with different number of latent dates $R = \{5, 7, 13, 15, 19, 25, 37\}$ and different number of latent spectral features $D' = \{4, 6, 9, 10, 11, 12, 13\}$. It can

---

[3]Experiments were also made with random irregular sampling and with selected dates from histogram of available dates. As modifying the positions of the latent dates do not have any influence on the performances, the simplest method was selected: regular sampling.
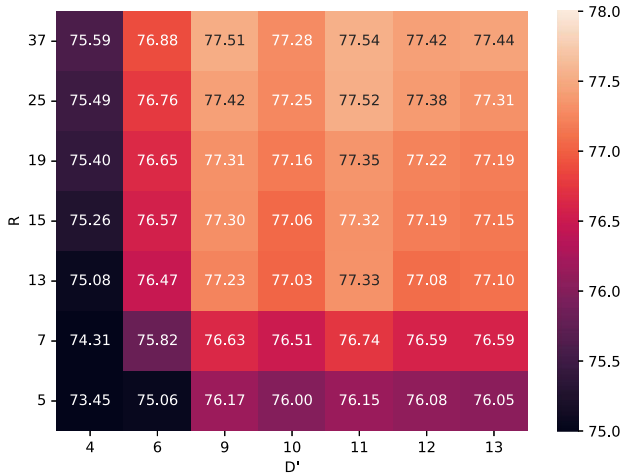
Fig. 4. Averaged overall accuracies (OA) for $H = 1$ (mean in % computed over nine different runs) with R the number of latent dates and D' the number of latent spectral features.
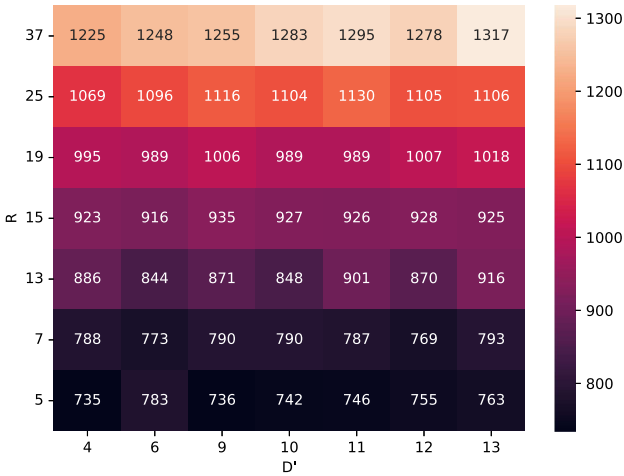

Fig. 5. Averaged training times in seconds for $H = 1$ (mean computed over nine different runs) with R the number of latent dates and D' the number of latent spectral features.

be seen that it is possible to reduce the number of parameters $\boldsymbol{\theta}_2$ by a factor four (i.e., from $584\,200$ to $165\,600$ parameters) and the training times by a factor two, while reaching the same OA. Indeed, as shown in Fig. 4, reducing R from 37 to 13 and D' from 13 to 9 has a negligible effect on the OA (i.e., from 77.44 to 77.23).

In addition, the number of heads H has a little impact on the classification performances as shown in Fig. 14 in Appendix B. Besides, from $H = 1$ to $H = 3$, the training time can be increased by a factor of two as shown in Fig. 15 in Appendix B.

C. Spatial Positional Encoding

The spatial information used to compute the positional encoded matrix $\mathbf{P}$ is composed of the spatial coordinates (northing ψ_1 and easting ψ_2) in meters in the Lambert 93 projection. The

TABLE V
AVERAGED OVERALL ACCURACIES (OA) WITH AND WITHOUT THE SPATIAL POSITIONAL ENCODED MATRIX $\mathbf{P}$ IN OUR ATTENTION-BASED INTERPOLATOR FOR THE MODEL *mTANe-SVGP* (MEAN % $\pm$ STANDARD DEVIATION COMPUTED WITH NINE RUNS)

	Without $\mathbf{P}$	With $\mathbf{P}$
OA	77.23 ± 0.17	78.63 ± 0.16

TABLE VI
AVERAGED OVERALL ACCURACIES (OA) (MEAN % $\pm$ STANDARD DEVIATION) AND AVERAGED TRAINING TIMES (IN SEC) FOR THE *mTANE-SVGP* WITH $R = 13$ LATENT DATES, $D' = 9$ LATENT SPECTRAL FEATURES, $H = 1$ HEAD AND THE SPATIAL POSITIONAL ENCODED MATRIX $\mathbf{P}$ FOR DIFFERENT NUMBER OF INDUCING POINTS M (COMPUTED OVER NINE RUNS)

	Number of inducing points M			
	50	100	150	200
Averaged OA	78.63 ± 0.16	79.20 ± 0.21	79.43 ± 0.29	79.48 ± 0.17
Training time	834	910	921	967

number of neurons in the first and second layers are, respectively, L_1 and L_2 and were found by trial and error: $L_1 = 16$ and $L_2 = 14$.

As shown in Table V, the use of the spatial positional encoding in the extended mTAN for the *mTANe-SVGP* model increased by nearly 1.5 points the OA. Besides, by using the spatial information through a spatial covariance function in [14], the OA was increased by nearly two points which is comparable to the results we obtained with the spatial positional encoding. The metrics per class for both models (without and with spatial positional encoding) are represented in Appendix C.

Fig. 6 represents the value of $\mathbf{P}$ for the features number 4 and number 12. This value was computed using different spatial coordinates on a regularly spaced grid over the 27 tiles. Fig. 6(a) and (b) exhibits smooth spatial transitions and anisotropric spatial similarity [see (9)].

D. Influence of the Number of Inducing Points

Fig. 7 represents the number of learnable parameters $\boldsymbol{\theta}_2$ based on the number of latent spectro–temporal features $R \times D'$ and the number of inducing points M. The reduction of the number of latent spectro–temporal from 481 ($R = 37, D' = 13$) to 117 ($R = 13, D' = 9$) results in a significant reduction of the number of learnable parameters $\boldsymbol{\theta}_2$ as shown in Fig. 7, with no loss in terms of classification accuracy (see previous discussion in Section IV-B). Furthermore, it is possible to double the number of inducing points from 50 to 100, while keeping the number of parameters $\boldsymbol{\theta}_2$ with $R = 13, D' = 9$ lower than with 50 inducing points and $R = 37, D' = 13$.

It is known that the learning capacity of SVGP is strongly influenced by the number of inducing points, and a tradeoff should be found between the computational complexity and the learning capacity [33]. By benefiting of a reduced computational load thanks to the dimension reduction, we perform several experiments with increasing number of inducing points $M = \{100, 150, 200\}$. Table VI represents averaged OAs and training times computed with different number of inducing points. With
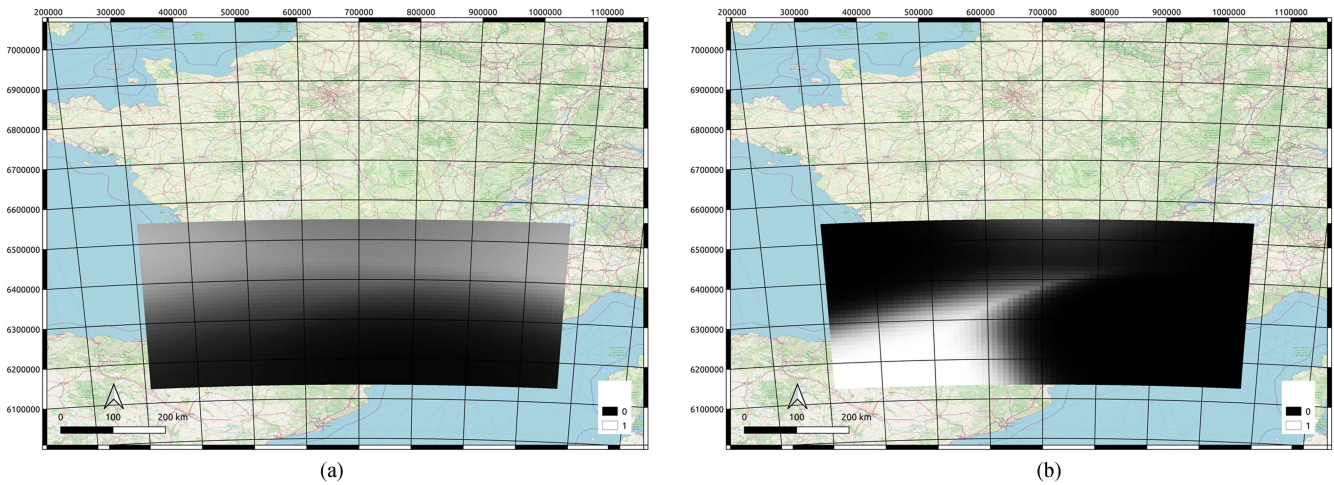
Fig. 6. Spatial positional encoding $\mathbf{P}$ computed over a regular grid of spatial coordinates (background map© OpenStreetMap contributors) (EPSG:2154). (a) Feature 4. (b) Feature 12.
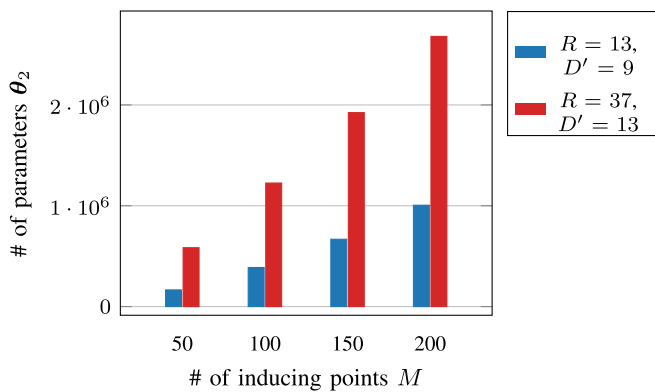

Fig. 7. Number of learnable parameters $\boldsymbol{\theta}_2$ based on the number of inducing points M and the number of spectro–temporal features $R \times D'$.

Fig. 8. Boxplots of the OA for each model (*mTANe-SVGP*, *mTANe-MLP*, *mTANe-LTAE,* and *raw-LTAE*) computed over nine runs.

$M = 200$, the OA is increased by almost one point compared to $M = 50$. Training time is only slightly affected by this increase in the number of inducing points, i.e., 834 s to 967 s. Hence, spectro–temporal reduction made possible to use higher number of inducing points and thus to increase the performances, while maintaining a reduced computational load.

V. COMPARISON WITH COMPETITIVE METHODS

This section presents a comparison of the *mTANe-SVGP* model with different models: *mTANe-MLP*, *mTANe-LTAE*, and *raw-LTAE*. First, the performance results are studied quantitatively and qualitatively.

In the following, from the results obtained in the previously in Section IV, the interpolator is set-up with $R = 13$ latent dates, $D' = 9$ latent spectral features, $H = 1$ head, spatial positional encoding matrix $\mathbf{P}$ and $M = 200$ inducing points. As the *raw-LTAE* model is the only one not using the interpolator, no spectral or temporal reduction was implemented in this model.
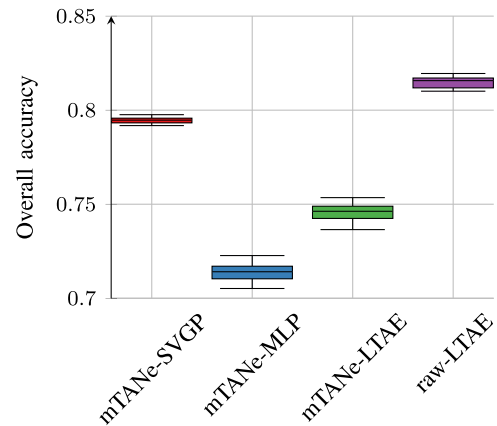
TABLE VII
AVERAGED TRAINING TIMES (IN SEC) COMPUTED OVER NINE RUNS AND
NUMBER OF TRAINABLE PARAMETERS FOR EACH MODEL (*MTANE-SVGP*,
MTANE-MLP, *MTANE-LTAE,* AND *RAW-LTAE*)

	mTANe-SVGP	mTANe-MLP	mTANe-LTAE	raw-LTAE
Training time	967	1207	840	1279
# parameters	1 005 675	33 113	184 376	761 380

A. Quantitative Results

Classification accuracies are given in Fig. 8. From the results, the SVGP model took greater advantage of the interpolator than the MLP or the LTAE models. Indeed, the OA of the *mTANe-SVGP* model is seven points above the *mTANe-MLP* model and around four points above the *mTANe-LTAE* model. On the other hand, the *mTANe-SVGP* model is in average two points below the *raw-LTAE* model. The F-score, recall, and precision per class for the *mTANe-SVGP*, *mTANe-MLP*, *mTANe-LTAE,* and *raw-LTAE* are represented in Figs. 16, 17 and 18 in Appendix C.

The number of trainable parameters and the training times for each method are summarized in Table VII. The *mTANe-SVGP* model has more trainable parameters that the *raw-LTAE* model. However, the training time of the *mTANe-SVGP* model is about 1.3 times shorter than the *raw-LTAE* as shown in Table VII. By using a spectro–temporal reduction with our interpolator, the number of trainable parameters for the *mTANe-SVGP* is just over 2.5 times lower than the simple SVGP (i.e., 1 005 675 versus 2 680 075), as described in Fig. 7. The number of parameters of the *raw-LTAE* is also very large because it is not able to deal with unaligned time series and therefore has to combine all the dates. By using our interpolator, for the LTAE, the number of trainable parameters is reduced by four. However, as shown in Fig. 8, the OA of the *mTANe-LTAE* model is seven points below the *raw-LTAE* model.

B. Qualitative Results

Land cover maps have been produced for each model (*mTANe-SVGP*, *mTANe-MLP*, *mTANe-LTAE*, *raw-LTAE*) using the iota^2 processing chain [34] on two different tiles: 31TCJ and 31TDJ. Inference was performed using the model trained on the 27 tiles with the best OA over the nine runs. The results obtained by, respectively, the *mTANe-SVGP* and *raw-LTAE* are shown in Fig. 9. The results obtained on this agricultural area on the 31TCJ tile showed that the main structures of the map are clearly represented (i.e., crop field border). Indeed, the classification map does not exhibit rounded borders as it is often the case with CNN model [35]. The *mTANe-SVGP* takes into account spatial information without spatial over-smoothing. Fig. 10 also represents land cover maps obtained with the *mTANe-SVGP* and *raw-LTAE* but in a mountainous area ("Montagne noire"). These maps clearly show some errors in predictions. Indeed, WAT class was predicted instead of GRA class for *raw-LTAE*. Moreover, STC and COR were predicted instead of WOM, for *mTANe-SVGP* and *raw-LTAE*, respectively. There are several reasons for these errors. First, there are no pixels from the *training* or *test* dataset in this area. Moreover, this area has a fairly high relief, at an altitude of 850 meters. All the generated land cover maps are available for download.[4]

VI. DISCUSSION

In this section, the latent representation and the similarity kernel learned by the interpolator are discussed. Moreover, an additional comparative study is made between the *mTANe-SVGP* and the *raw-LTAE* to evaluate the temporal sampling robustness.

A. Latent Representation

It is possible to visualize the learned latent representation $\hat{x}_j$. Fig. 11 represents the comparison of three NDVI time series profiles from one pixel labeled as "CORN": the raw data, the

(a)

(b)

(c)

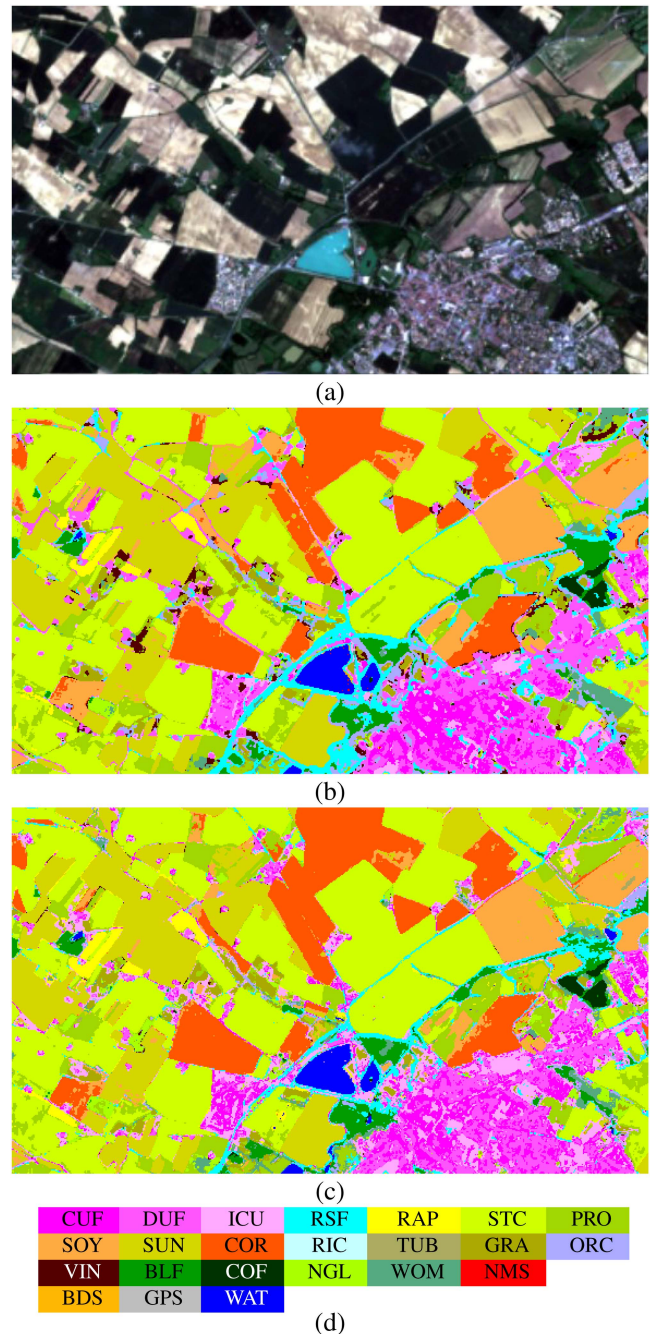| CUF	DUF	ICU	RSF	RAP	STC	PRO
SOY	SUN	COR	RIC	TUB	GRA	ORC
VIN	BLF	COF	NGL	WOM	NMS	
BDS	GPS	WAT				

(d)

Fig. 9. Land covers maps in an agricultural area. (a) Sentinel-2 true color composition, (b) Classification map obtained with *mTANe-SVGP*, (c) Classification map obtained with *raw-LTAE* and (d) the class color map (see Table III for correspondence).

gapfilled data (i.e., linearly interpolated) and the learned latent representation obtained by our time and space informed kernel interpolator.

The latent representation obtained in Fig. 11 clearly does not minimize the reconstruction error of the original time series. For instance, the second minimum of the NDVI observed around the day of the year 280 is not reconstructed. Yet, this is the representation that minimizes the classification loss function of the SVGP.
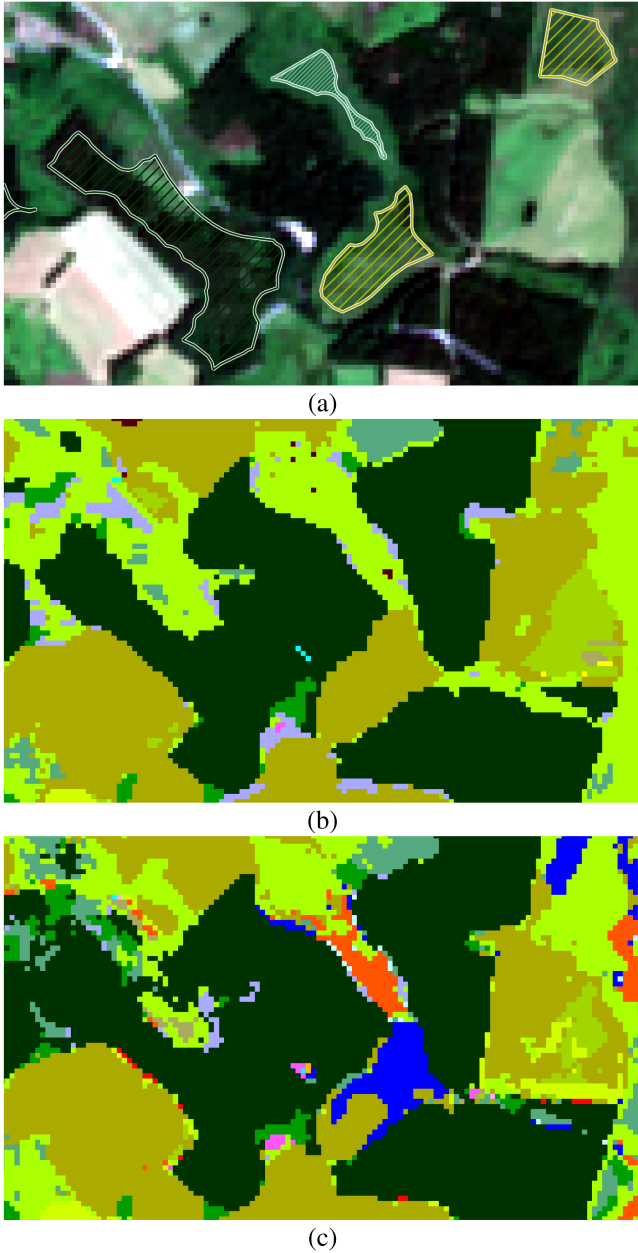
(a)

(b)

(c)

Fig. 10. Land covers maps in an agricultural area. (a) Sentinel-2 true color composition with referenced polygons, (b) Classification map obtained with *mTANe-SVGP* and (c) Classification map obtained with *raw-LTAE*. The class color map is represented in Fig. 9 (see Table III for correspondence).

B. Versatility of the Similarity Kernel

As defined in Section II-A3, by using *attention* and *embedding* mechanisms, the similarity kernel is able to adapt to the pixel temporal sampling. The versatility of the similarity kernel can be shown by computing the attention value γ_{r_l} defined in (5) for different latent dates r_l and for different sets of observed dates $\mathbf{T}$. In Fig. 12, two different latent dates are studied $r_l = 181$ and $r_l = 361$.[5] For each latent date r_l, two different sets of observed dates $\mathbf{T}$ are considered. First, the attention value was computed with a regular set of observed dates: $\mathbf{T} = \{1, \ldots, 365\}$ with an

[5]The attention value plotted was normalized (cf Fig. 12) in order to have the same vertical scale.
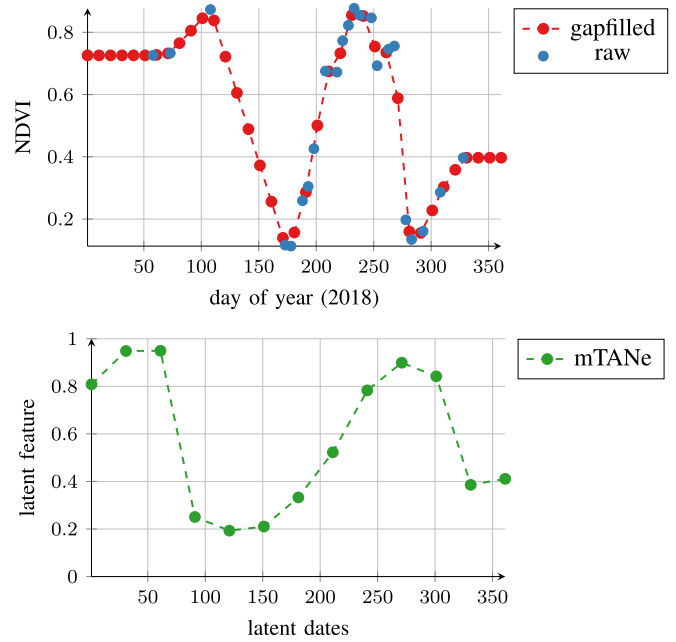

Fig. 11. NDVI time series profiles for a pixel labeled "CORN". Blue points ● correspond to the raw data, the outlier values have been removed in order to have a comprehensive plot. Red points ● correspond to the value obtained with a linear interpolation with an interval of 10 days for a total of 37 dates. Green points ● correspond to the latent representation $\hat{\mathbf{x}}_j$ with j = NDVI obtained from the *mTANe-SVGP* model, before the spectral reduction ($D' = 9$).
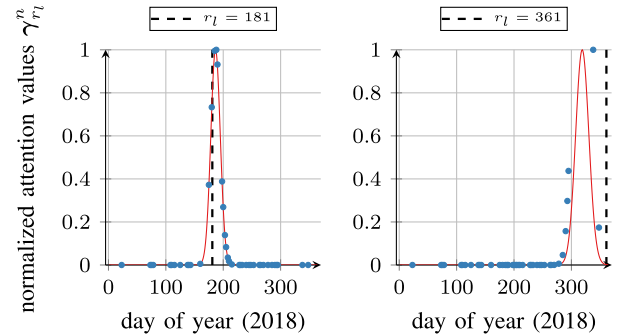

Fig. 12. Normalized attention values $\gamma_{r_l}^n = \frac{\gamma_{r_l}}{\max(\gamma_{r_l})}$ computed on two different latent dates $r_l = 181$ and $r_l = 361$. — corresponds to $\gamma_{r_l}^n$ computed with $\mathbf{T} = \{1, \ldots, 365\}$ with a regular interval of $\tau = 1$ day. Blue points ● correspond to $\gamma_{r_l}^n$ computed with $\mathbf{T} = \mathbf{T}^i$ for a random pixel i.

interval of $\tau = 1$ day (in red in Fig. 12). Then, the attention value was computed with a random set of observed dates from a pixel i with $\mathbf{T} = \mathbf{T}^i$ (in blue in Fig. 12).

From Fig. 12, we can see that contrary to conventional RBF kernel, the learned kernel is not centered on the latent date. It thus adapts itself according to the latent date r_l and the available observations. Moreover, for the set of observed dates $\mathbf{T} = \{1, \ldots, 365\}$ (i.e., continuous red line), the bandwidth is larger for the latent date $r_l = 361$ than for the latent date $r_l = 181$. Such property is referred to as a *variable-bandwidth* kernel in the statistical literature [36]. While it has shown to perform well on several cases, such kernel was difficult to optimize with standard statistical models. Using the proposed framework, the optimization is efficient, scales well and can handle any timestamp.
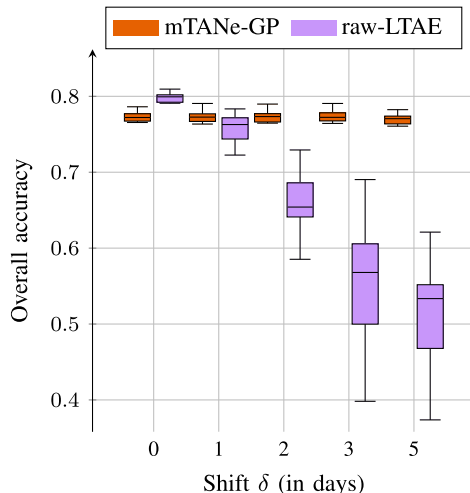
Fig. 13. Boxplots of the OA for the *mTANe-SVGP* and *raw-LTAE* models computed with the *test* data set only on the 31TCJ tile over nine runs. The models were trained and validated on the all 27 tiles. The acquisition dates **T** for the *test* data set were artificially shifted with different values: $\delta = \{0, 1, 2, 3, 5\}$ days.

C. Robustness to the Temporal Sampling

In Section V-A, the *raw-LTAE* showed better classification performances. However, to compute the inference on a specific area (e.g., on a specific Sentinel-2 tile), the *raw-LTAE* required having seen the whole set of observed dates during the training step. Thus, each possible temporal feature should have been encoded by the *raw-LTAE* during the training step in order to be able to classify a given area at inference time. This is not the case for our proposed model which is able to process pixels with any set of observed dates.

This can make the time encoding of *raw-LTAE* not robust to variations of the temporal sampling between the train and test sets, with a possible overfit on the training dates. To investigate this possible issue, dates not seen during the training step and used only for the inference were artificially created. They correspond to the original acquisition dates **T** from the *training* dataset that have been slightly shifted for the *test* data set. Different values for the shift were studied: $\delta = \{0, 1, 2, 3, 5\}$ days. Five days correspond to the maximum number of days between acquisition dates for pixels on two adjacent orbits. The OA was computed only on test samples from 31TCJ tile for two models *mTANe-SVGP* and *raw-LTAE* both trained on the 27 tiles.

From Fig. 13, we can say that the OA of the *mTANe-SVGP* model is not affected by this temporal shift δ. However, the OA of the *raw-LTAE* model is drastically impacted by the temporal shift δ. For a shit of one day, the OA is reduced by almost 3 points and it is almost divided by 1.5 with $\delta = 5$ days. The use of a time and space informed kernel interpolator makes the *mTANe-SVGP* model more robust to this shift than the *raw-LTAE* model which uses spectro–temporal attention mechanisms but no interpolation. We conclude that the *raw-LTAE* is more sensitive to dates seen during the training step and may therefore be likely to over-fit.

VII. CONCLUSION

This work introduces an approach to classify massive irregular and unaligned Sentinel-2 SITS. To deal with irregular and unaligned pixel time series, an end-to-end interpolation and learning strategy is proposed. A first module, a time and space informed kernel interpolator, enables to map irregular and unaligned SITS onto a fixed and reduced size latent space. Temporal reconstruction and spectral reduction were performed jointly but independently. This constrained spectro–temporal structure enables to reduce the complexity of the classifier. Indeed, the representation obtained is given to a SVGP classifier and all the parameters are jointly optimized during the optimization of the classifier. The spatial information is taken into account in the learned representation through a *spatial positional encoding*. Experiments were conducted on 27 Sentinel-2 tiles of the full year 2018 in the south of France.

In terms of accuracy, the end-to-end learning *mTANe-SVGP* model outperformed the simple SVGP classifier with linearly interpolated data (*Gapfilled-SVGP*). The significant reduction for the spectro-temporal features has allowed to use more inducing points while keeping the same complexity, resulting in improved classification performance. Moreover, the *mTANe-SVGP* model is above the *mTANe-MLP* and *mTANe-LTAE* models in terms of accuracy. While the proposed *mTANe-SVGP* does not outperform the *raw-LTAE* model, our model does not require for the inference the common temporal grid used during the training step. Besides, our result showed that the *raw-LTAE* model is sensitive to the set of available dates during inference, contrary to the proposed *mTANe-SVGP*, which showed stable performances. The *mTANe-SVGP* is therefore more likely to generalize well to large-scale scenario where irregular and variable sampling dates are prominent.

In this article, the potential of the multihead attention has not been fully taken into account. Indeed, only one head was used $H = 1$ and the performances with an increasing number of heads were not satisfying. A perspective of this work could be to inform the different heads with the spatial information: the linear layer $\boldsymbol{\beta}_H$ in (6) could be replaced by the output of a perceptron using the *spatial positional encoding*. This could help the heads to specialize spatially and differentiate themselves.

Another perspective of this work is to combine multimodal time series. Adding a radar sensor (i.e., Sentinel-1) or other type of optical sensors (i.e., Landsat 8 with its thermal bands) could improve the representation for the classification task. The ability of the interpolator to process unaligned time series would make the fusion of multisensor data straightforward. Moreover, in addition to spatial data (i.e., longitude and latitude), topographic data can be used to construct the *spatial positional encoding* in order to take better account of climatic, geographical, and other differences.

In the interest of reproducible research, the implementation of all the models (*mTANe-GP*, *mTANe-MLP*, *mTANe-LTAE* and *raw-LTAE*) is made available in the following repository:https://gitlab.com/Valentine-Bellet/land_cover_southfrance_mtan_gp_irregular_sits.

APPENDIX A
SOLVER PARAMETERS FOR EACH MODEL

TABLE VIII
PARAMETER VALUES FOR THE ADAM OPTIMIZER FOR THE MODELS: *GAPFILLED-SVGP*, *MTANE-SVGP*, *MTANE-MLP*, *MTANE-LTAE,* AND *RAW-LTAE*

	Gapfilled-SVGP	mTANe-SVGP	mTANe-MLP	mTANe-LTAE	raw-LTAE
Number of epochs	100	100	300	100	100
Batch size	1024	1024	1000	1000	1000
Learning rate	1×10^{-3}	1×10^{-3}	1×10^{-4}	5×10^{-5}	1×10^{-4}

APPENDIX B
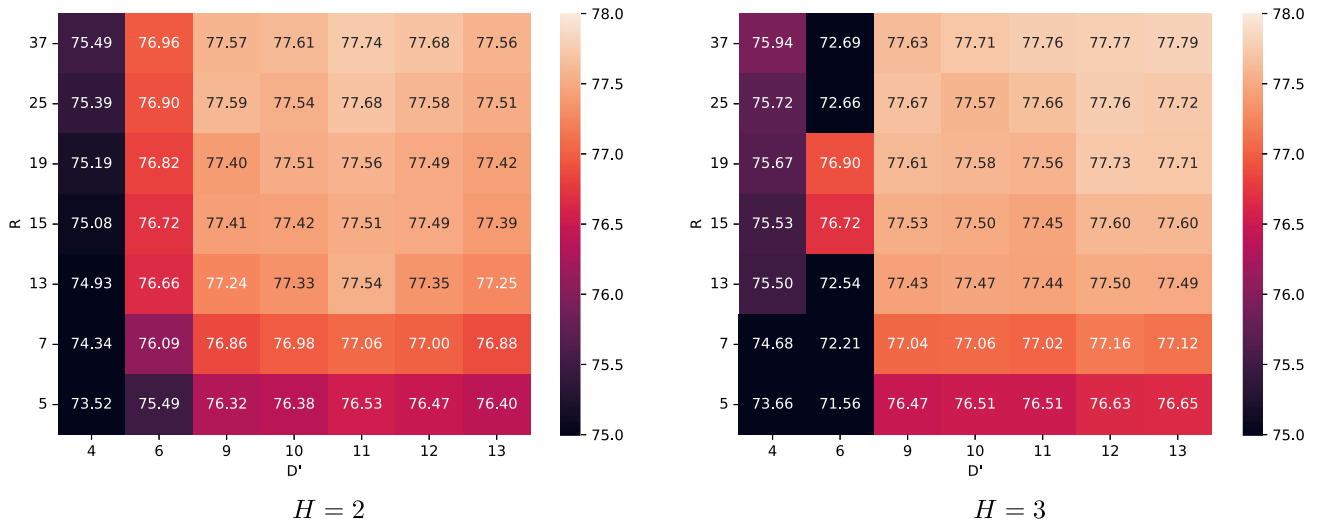## INFLUENCE OF THE SPECTRAL AND TEMPORAL REDUCTION FOR DIFFERENT H HEADS

Fig. 14.   Averaged overall accuracies (OA) (mean in % computed over nine different runs) with R the number of latent dates, D' the number of latent spectral features and H the number of heads.
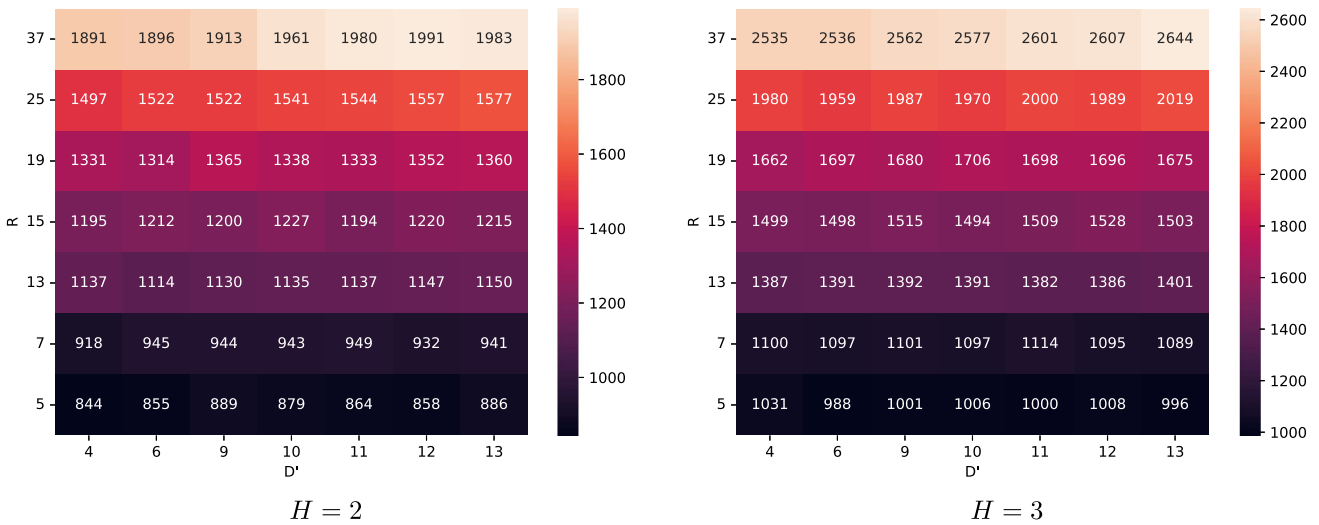

Fig. 15.   Averaged training times in seconds (mean computed over nine different runs) with R the number of latent dates, D' the number of latent spectral features and H the number of heads.
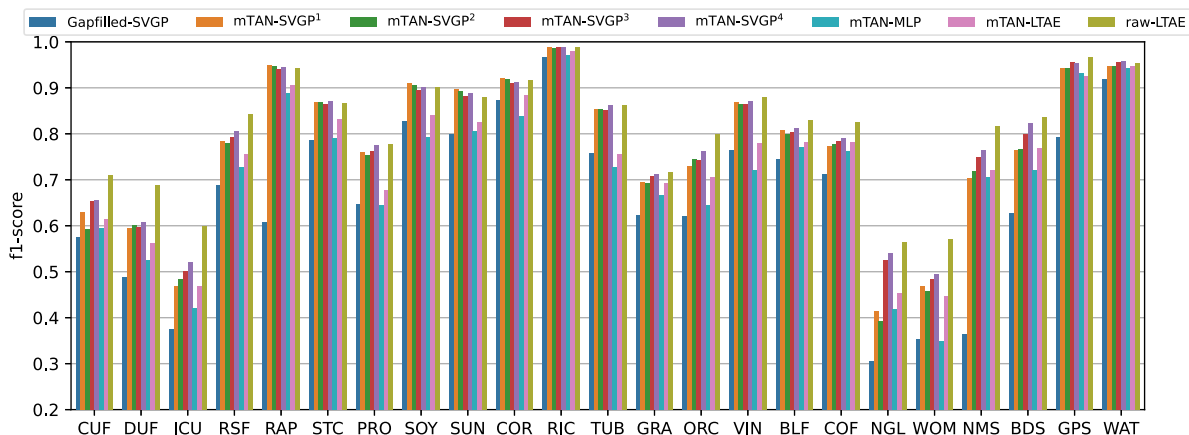
APPENDIX C
METRICS PER CLASS

Fig. 16.　Averaged F-score per class. mTANe-SVGP[1]: $H = 1, D' = 13, R = 37, M = 50$; mTANe-SVGP[2]: $H = 1, D' = 9, R = 13, M = 50$; mTANe-SVGP[3]: with $\mathbf{P}$ and $H = 1, D' = 9, R = 13, M = 50$; mTANe-SVGP[4]: with $\mathbf{P}$ and $H = 1, D' = 9, R = 13, M = 200$.

Fig. 17.　Averaged recall per class. mTANe-SVGP[1]: $H = 1, D' = 13, R = 37, M = 50$; mTANe-SVGP[2]: $H = 1, D' = 9, R = 13, M = 50$; mTANe-SVGP[3]: with $\mathbf{P}$ and $H = 1, D' = 9, R = 13, M = 50$; mTANe-SVGP[4]: with $\mathbf{P}$ and $H = 1, D' = 9, R = 13, M = 200$.
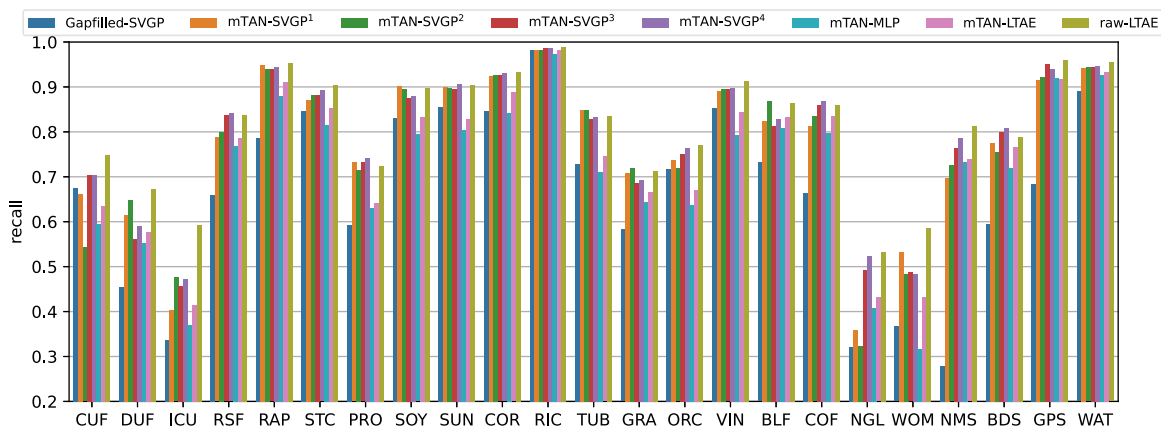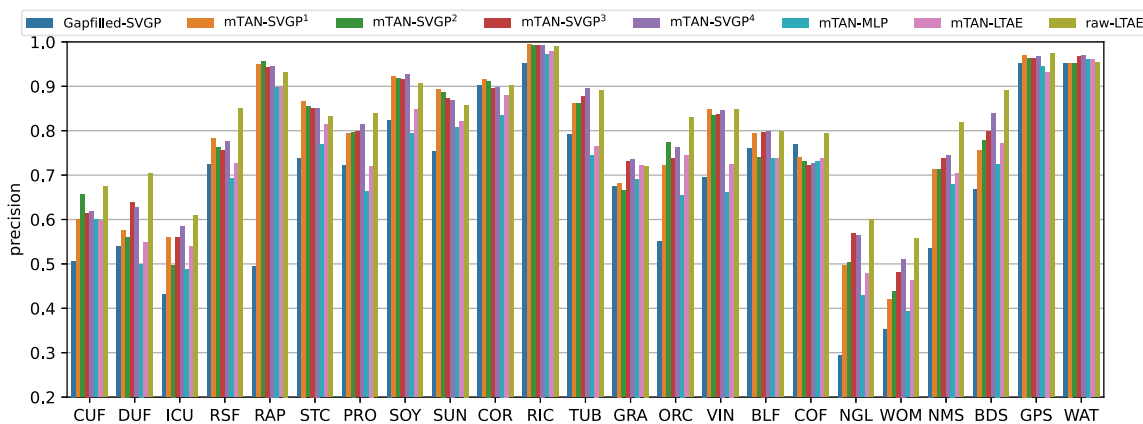

Fig. 18.　Averaged precision per class. mTANe-SVGP[1]: $H = 1, D' = 13, R = 37, M = 50$; mTANe-SVGP[2]: $H = 1, D' = 9, R = 13, M = 50$; mTANe-SVGP[3]: with $\mathbf{P}$ and $H = 1, D' = 9, R = 13, M = 50$; mTANe-SVGP[4]: with $\mathbf{P}$ and $H = 1, D' = 9, R = 13, M = 200$.

REFERENCES

[1] IPCC, "Synthesis report of the IPCC sixth assessment report (ar6) longer report," 2023. [Online]. Available: https://report.ipcc.ch/ar6syr/pdf/IPCC_AR6_SYR_LongerReport.pdf

[2] C. Persello et al., "Deep learning and earth observation to support the sustainable development goals: Current approaches, open challenges, and future opportunities," *IEEE Geosci. Remote Sens. Mag.*, vol. 10, no. 2, pp. 172–200, Jun. 2022.

[3] D. Tuia et al., "Artificial intelligence to advance earth observation: A perspective," 2023, *arXiv:2305.08413*.

[4] F. Bertini et al., "Sentinel-2 esa's optical high-resolution mission for gmes operational services," *ESA Bull. Bull. ASE. Eur. Space Agency*, vol. SP-1322, pp. 9–13, 2012.

[5] G. Camps-Valls, "Machine learning in remote sensing data processing," in *Proc. IEEE Int. Workshop Mach. Learn. Signal Process.*, pp. 1–6, 2009.

[6] G. Camps-Valls, D. Tuia, L. Bruzzone, and J. A. Benediktsson, "Advances in hyperspectral image classification: Earth monitoring with statistical learning methods," *IEEE Signal Process. Mag.*, vol. 31, no. 1, pp. 45–54, Jan. 2014.

[7] X. X. Zhu et al., "Deep learning in remote sensing: A comprehensive review and list of resources," *IEEE Geosci. Remote Sens. Mag.*, vol. 5, no. 4, pp. 8–36, Dec. 2017.

[8] M. Belgiu and L. Drăguţ, "Random forest in remote sensing: A review of applications and future directions," *ISPRS J. Photogrammetry Remote Sens.*, vol. 114, pp. 24–31, Apr. 2016.

[9] J. Inglada, A. Vincent, M. Arias, B. Tardy, D. Morin, and I. Rodes, "Operational high resolution land cover map production at the country scale using satellite image time series," *Remote Sens.*, vol. 9, no. 1, 2017, Art. no. 95.

[10] C. Pelletier, S. Valero, J. Inglada, N. Champion, and G. Dedieu, "Assessing the robustness of random forests to map land cover with high resolution satellite image time series over large areas," *Remote Sens. Environ.*, vol. 187, pp. 156–168, 2016.

[11] M. Rußwurm and M. Körner, "Multi-temporal land cover classification with sequential recurrent encoders," *ISPRS Int. J. Geo- Inf.*, vol. 7, no. 4, 2018, Art. no. 129.

[12] C. Pelletier, G. Webb, and F. Petitjean, "Temporal convolutional neural network for the classification of satellite image time series," *Remote Sens.*, vol. 11, Mar. 2019, Art. no. 523.

[13] V. S. F. Garnot and L. Landrieu, "Lightweight temporal self-attention for classifying satellite images time series," in *Proc. Int. Workshop Adv. Anal. Learn. Temporal Data*, AALTD, . 2020, pp. 171–181.

[14] V. Bellet, M. Fauvel, and J. Inglada, "Land cover classification with Gaussian processes using spatio-spectro-temporal features," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 4400221.

[15] S. C.-X. Li and B. Marlin, "A scalable end-to-end gaussian process adapter for irregularly sampled time series classification," in *Proc. 30th Int. Conf. Neural Inf. Process. Syst.*, 2016, pp. 1812–1820.

[16] A. Constantin, M. Fauvel, and S. Girard, "Joint supervised classification and reconstruction of irregularly sampled satellite image times series," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 4403913.

[17] F. Petitjean, J. Inglada, and P. Gancarski, "Satellite image time series analysis under time warping," *IEEE Trans. Geosci. Remote Sens.*, vol. 50, no. 8, pp. 3081–3095, Aug. 2012.

[18] V. Maus, G. Câmara, M. Appel, and E. Pebesma, "dtwsat: Time-weighted dynamic time warping for satellite image time series analysis in R," *J. Stat. Softw.*, vol. 88, no. 5, pp. 1–31, 2019.

[19] V. Maus, G. Câmara, R. Cartaxo, A. Sanchez, F. M. Ramos, and G. R. de Queiroz, "A time-weighted dynamic time warping method for land-use and land-cover mapping," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 9, no. 8, pp. 3729–3739, Aug. 2016.

[20] S. S. T. de Oliveira, L. M. L. Pascoal, M. d. C. Cardoso, E. F. Bueno, V. J. S. Rodrigues, and W. S. Martins, "A parallel and distributed approach to the analysis of time series on remote sensing Big Data," *J. Inf. Data Manage.*, vol. 10, pp. 16–34, Jun. 2019.

[21] D. Neil, M. Pfeiffer, and S.-C. Liu, "Phased LSTM: Accelerating recurrent network training for long or event-based sequences," in *Proc. 30th Int. Conf. Neural Inf. Process. Syst.*, 2016, pp. 3889–3897.

[22] D. Ienco, R. Gaetano, C. Dupaquier, and P. Maurel, "Land cover classification via multitemporal spatial data by deep recurrent neural networks," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 10, pp. 1685–1689, Oct. 2017.

[23] A. Vaswani et al., "Attention is all you need," in *Proc. 31st Int. Conf. Adv. Neural Inf. Process. Syst.*, 2017, vol. 30, pp. 6000–6010.

[24] M. Rußwurm and M. Körner, "Self-attention for raw optical satellite time series classification," *ISPRS J. Photogrammetry Remote Sens.*, vol. 169, pp. 421–435, 2020.

[25] S. N. Shukla and B. M. Marlin, "Interpolation-prediction networks for irregularly sampled time series," in *Proc. 7th Int. Conf. Learn. Representations*, New Orleans, LA, USA, 2019.

[26] S. N. Shukla and B. Marlin, "Multi-time attention networks for irregularly sampled time series," in *Proc. Int. Conf. Learn. Representations*, 2021.

[27] T. Hastie, R. Tibshirani, and J. Friedman, "The elements of statistical learning," in *Springer Series in Statistics*. New York, NY, USA: Springer, 2001.

[28] A. Zhang, Z. C. Lipton, M. Li, and A. J. Smola, "Dive into deep learning," 2021, *arXiv:2106.11342*.

[29] L. Baudoux, J. Inglada, and C. Mallet, "Toward a yearly country-scale CORINE land-cover map without using images: A map translation approach," *Remote Sens.*, vol. 13, Mar. 2021, Art. no. 1060.

[30] L. Baetens, C. Desjardins, and O. Hagolle, "Validation of copernicus Sentinel-2 cloud masks obtained from MAJA, Sen2Cor, and FMask processors using reference cloud masks generated with a supervised active learning procedure," *Remote Sens.*, vol. 11, Feb. 2019, Art. no. 433.

[31] O. D. Team, "Orfeo ToolBox 7.1," Mar. 2020. [Online]. Available: https://zenodo.org/record/3715021

[32] M. Kuhn and K. Johnson, *Feature Engineering and Selection: A Practical Approach for Predictive Models*, 1 ed. London, U.K.: Chapman & Hall/CRC, Jul. 2019.

[33] J. Hensman, A. Matthews, and Z. Ghahramani, "Scalable variational gaussian process classification," in *Proc. 18th Int. Conf. Artif. Intell. Statist.*, 2015, pp. 351–360.

[34] J. Inglada, A. Vincent, M. Arias, and B. Tardy, "IoTA2-a25386," Zenodo, Jul. 17, 2016, doi: 10.5281/zenodo.58150.

[35] A. Stoian, V. Poulain, J. Inglada, V. Poughon, and D. Derksen, "Land cover maps production with high resolution satellite image time series and convolutional neural networks: Adaptations and limits for operational systems," *Remote Sens.*, vol. 11, Aug. 2019, Art. no. 1986.

[36] G. R. Terrell and D. W. Scott, "Variable kernel density estimation," *Ann. Statist.*, vol. 20, no. 3, pp. 1236–1265, 1992.

Valentine Bellet (Student Member, IEEE) received the master's degree in automatic control and electronics from INSA Toulouse, Toulouse, France, in 2019. She is currently working toward the Ph.D. degree in artificial intelligence for ecosystem monitoring using satellite image time series with the Centre d'Études Spatiales de la Biosphère (CESBIO) Laboratory, Université de Toulouse, Toulouse, France.

She is working on the subject of land cover pixel-based classification with satellite image time series (SITS) at a national scale.

Mathieu Fauvel (Senior Member, IEEE) received the Ph.D. degree in image and signal processing from the Grenoble Institute of Technology, Grenoble, France, in 2007.

From 2008 to 2010, he was a post-doctoral researcher with the MISTIS Team, National Institute for Research in Digital Science and Technology (INRIA). From 2011 to 2018, he was an Associate Professor with the DYNAFOR Lab (INRA), National Polytechnic Institute of Toulouse, Toulouse, France. Since 2018, he has been a Researcher with INRAe and Centre d'Études Spatiales de la Biosphère (CESBIO) Laboratory, Université de Toulouse, Toulouse, France. His research interests are remote sensing, machine learning, and image processing.

Jordi Inglada received the master's degree in telecommunications engineering from the Universitat Politècnica de Catalunya, Barcelona, Spain, and the École Nationale Supérieure des Télécommunications de Bretagne, Brest, France, in 1997, and the Ph.D. degree in signal processing and telecommunications from the Université de Rennes 1, Rennes, France, in 2000.

He is currently with the Centre National d'Études Spatiales (French Space Agency), Toulouse, France, where he is involved in the field of remote sensing image processing at the Centre d'Études Spatiales de la Biosphère (CESBIO) Laboratory. He is involved in the development of image processing algorithms for the operational exploitation of Earth observation images, mainly in the field of multitemporal image analysis for land use and cover change.

Julien Michel received the Telecommunications Engineer degree from the École Nationale Supérieure des Télécommunications de Bretagne, Brest, France, in 2006. He is currently working toward the Ph.D. degree in machine learning methods for the improvement of spatial and temporal resolutions in Satellite Images Time-Series, and application in the Thermal Infra-Red domain with the Centre d'Etudes Spatiales de la Biosphère (CESBIO), Université de Toulouse, Toulouse, France.

He is currently with the Centre National d'Etudes Spatiales (French Space Agency), Université de Toulouse, Toulouse, France, where he is working as a Research Engineer on remote sensing image processing with the Centre d'Etudes Spatiales de la Biosphère (CESBIO). His main research topic focuses on the fusion of heterogeneous Satellite Image Time Series (SITS) so as to enhance their spatial and temporal resolutions. His wider research interests include image processing and machine learning for remote sensing data.