



# 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, Mathieu Fauvel, Jordi Inglada, Julien Michel

## ► To cite this version:

Valentine Bellet, Mathieu Fauvel, Jordi Inglada, Julien Michel. End-to-end Learning for Land Cover Classification using Irregular and Unaligned SITS by Combining Attention-Based Interpolation with Sparse Variational Gaussian Processes. 2023. hal-04112115v3

**HAL Id: hal-04112115**

**<https://hal.science/hal-04112115v3>**

Preprint submitted on 13 Jul 2023 (v3), last revised 21 Dec 2023 (v4)

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# 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, *Graduate 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 an attention-based interpolator: the Multi-Time Attention Networks (mTAN), with a Sparse Variational Gaussian Processes (SVGP) classifier. The mTAN projects irregular and unaligned SITS onto a fixed and reduced temporal grid latent representation. The resulting structured feature representation takes into account the spectro-temporal correlation of the SITS. Moreover, the spatial information is added to the interpolator 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<sup>2</sup> (about two billion pixels) in the south of France. 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 deep learning classifiers (respectively seven and four points for the Multi-layer Perceptron and the Lightweight Temporal Self-Attention classifiers).

**Index Terms**—Satellite Image Time-Series (SITS), Sentinel-2, Land Cover Map, Pixel-Based, Classification, Large Scale, Sparse Variational Gaussian Processes, Earth Observation (EO), Remote Sensing, Representation Learning.

## I. INTRODUCTION

IN March 2023, the final synthesis report of the Sixth Assessment Report (AR6) was released by the Intergovernmental Panel on Climate Change (IPCC). 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 10m, six at 20m and three at 60m 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]–[7].

A widely used ML algorithm for pixel-wise classification, with very good performances even in large scale, is the Random Forest (RF) [8]–[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, in a previous work [14], 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. Yet this method, like most

This work is supported by the Natural Intelligence Toulouse Institute (ANITI) from Université Fédérale Toulouse Midi-Pyrénées under grant agreement ANITI ANR-19-PI3A-0004 (this PhD is co-founded by CS-Group and by the Centre National d'Études Spatiales (CNES)).

V. Bellet, M. Fauvel, J. Inglada and J. Michel are with CESBIO, Université de Toulouse, CNES/CNRS/INRAe/IRD/UPS, 31000 Toulouse, France (e-mail: valentine.bellet@univ-toulouse.fr, mathieu.fauvel@inrae.fr, jordi.inglada@cesbio.eu, julien.michel4@univ-tlse3.fr)



of the methods in literature, requires data with a fixed-size i.e. the number of features per pixel is the same for each pixel of the data to be processed.

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. In this context, 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 Random Forest (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 intra-annual 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, the Spatial Parallel TWDTW allowed to parallelize the TWDTW algorithm and to take into account the spatial dimension [20]. Even if it achieved almost linear speed up, it was not able to deal with very large data-sets.

Few DL methods can directly deal with these irregular and unaligned time series. For example, Long Short-Term Memory (LSTM) [21] can take into account irregular time series, however, they do not support unaligned time series and are slow to train because of the lack of parallelization abilities. Transformer architectures [22], 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 [23] 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.

In order to profit from the advantages of the above-mentioned SVGP approach [14], in this work, we propose to learn a fixed-size latent representation as a pre-processing step to the classifier. Shukla *et al.* [24] designed a kernel smoother to build representations for irregular time series. Recently, the authors of [25] proposed a method called Multi-Time Attention

Networks (mTAN) which enables working with irregular and unaligned time series. mTAN produces a fixed representation by using multiple continuous time embeddings coupled with attention mechanisms. By using end-to-end training (mTAN coupled with a classifier), the performance results were similar to or better than the state-of-the-art classifiers [25]. In this article, the mTAN is adapted to project the irregular and unaligned time series onto a latent space of fixed and reduced size. The obtained representation is then given to the SVGP classifier and all the parameters are jointly optimized using the loss function associated to SVGP.

In large scale classification, due to different climatic and topographic conditions, there is a variation of the spectro-temporal signature over the spatial domain (i.e. non stationarity). By using spatial coordinates, with spatial stratification [9] or by learning a spatial-informed classifier [14], this non-stationarity can be taken into account and performances are improved. In this work, we propose to add the spatial information before the classification, in the latent representation obtained by the mTAN. The method used is based on the *spatial positional encoding* as defined in [26].

In a previous work [14], we found that the SVGP performs well but its complexity is related to the number of spectro-temporal variables and their inter-correlation. Therefore, we propose to learn the representation with a reduced temporal grid and with a low number of spectral features in order to reduce the complexity of the SVGP and simplify its optimization. Besides, the number of inducing points (i.e. parameters of the SVGP classifier) has a strong influence on the quality of the approximation and therefore on the classification performance. Thus, we propose to study the influence of the number of inducing points. w.r.t. the latent space size.

The remainder of this paper is organized as follows. Section II describes how the mTAN is used to process irregular and unaligned pixel time series. Section II-B defines how the mTAN is extended in order to fit to the classification task in particular by using spectro-temporal reduction and spatial positional encoding. The experimental setup is detailed in Section III. The results obtained with the end-to-end trained model (mTAN coupled with SVGP) are provided in Section IV. Different competitive methods are studied and their associated results are presented in Section V. Finally, Section VI concludes this paper and opens discussions on future works.

## II. METHODS

### A. Attention-based temporal interpolator

This section describes how irregular and unaligned pixel time series are projected onto a fixed temporal grid in order to be used by the classifier. First, some notations and definitions which are used throughout this paper are introduced. Then, the latent interpolator at the core of the proposed method is presented and its modification is described in the last part.

1) *Notations and definitions:* In this paper, the  $i$ th pixel time series  $\mathbf{x}^i(t_k)$  at time  $t_k$  is defined by its spectral measurements  $\{x_1^i(t_k), \dots, x_j^i(t_k), \dots, x_D^i(t_k)\}$  with  $i \in \{1, \dots, N\}$ ,  $N$  the number of pixels and  $D$  the number of spectral features.

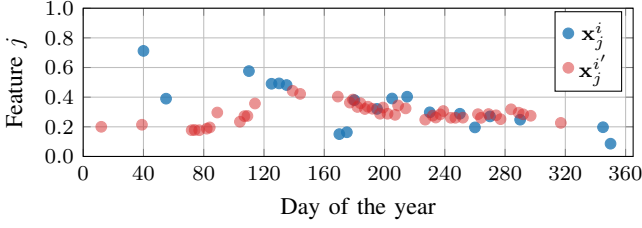


Fig. 1:  $\mathbf{x}_j^i$  and  $\mathbf{x}_j^{i'}$  are two irregular and unaligned time series for pixels  $i$  and  $i'$  respectively, for the spectral feature  $j$ .

Additionally, two spatial coordinates  $\psi_1^i$  and  $\psi_2^i$  are associated to the pixel  $\mathbf{x}^i$ . Moreover,  $y^i \in \{1, \dots, C\}$  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:  $\mathbf{T}_j^i = \{t_{j1}^i, \dots, t_{jk}^i, \dots, t_{jT_j^i}^i\}$ , 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 multi-source 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:

$$\begin{aligned} \mathbf{T} &= \bigcup_{i=1}^N \mathbf{T}^i \\ &= \{t_1, \dots, t_k, \dots, t_T\} \end{aligned}$$

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

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

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

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

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:  $\mathbf{R} = \{r_1, \dots, r_l, \dots, r_R\}$ . In this work, we focus on the well-established Nadaraya-Watson kernel smoother [27, Chapter 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<sup>1</sup>:

$$\hat{x}_j(r_l) = \frac{\sum_{t_k=t_1}^{t_T} K(r_l, t_k) m(t_k) x_j^*(t_k)}{\sum_{t'_k=t_1}^{t_T} K(r_l, t'_k) m(t'_k)} \quad (3)$$

with  $K$  some similarity kernel [27, Chapter 6]. Usually, the 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  $\sigma$ , 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, Chapter 11]. In the following, the Multi Time Attention Networks (mTAN) [25] is discussed as an extension of the kernel smoother to build the interpolator for the classification model in our end-to-end training.

3) *Multi Time Attention Networks (mTAN)*: To build the interpolator, Shukla *et al.* [25] proposed using attention mechanisms and more precisely the scaled-dot product attention defined in [22].

Firstly, a learnable time embedding function (named *temporal positional encoding*)  $\phi$  is defined. It maps a given  $t$  onto a higher dimensional space of size  $E$  such as:

$$\begin{aligned} \phi: \mathbb{R} &\rightarrow \mathbb{R}^E \\ t &\mapsto \phi(t) = \begin{bmatrix} \omega_1 t + \alpha_1 \\ \sin(\omega_2 t + \alpha_2) \\ \vdots \\ \sin(\omega_E t + \alpha_E) \end{bmatrix} \end{aligned} \quad (4)$$

with  $\omega_p$  and  $\alpha_p$ ,  $p \in \{1, \dots, E\}$ , the learnable parameters.

Then, this embedding  $\phi$  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 [22].

Finally, (3) can be re-written using a masked softmax operator [28, Chapter 11.3.2] such as:

$$\begin{aligned} \hat{x}_j(r_l) &= \text{softmax} \left\{ \frac{(\Phi(\mathbf{T})^\top \mathbf{W}_k^\top \mathbf{W}_q \phi(r_l)) \odot \mathbf{m}}{\sqrt{E}} \right\}^\top \mathbf{x}_j^* \\ &= \gamma_{r_l}^\top \mathbf{x}_j^*. \end{aligned} \quad (5)$$

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 [22].

<sup>1</sup>For clarity, we consider only one pixel and we drop the index  $i$  in the remaining of the paper.

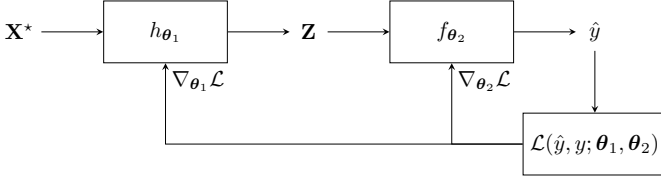


Fig. 2: End-to-end learning for the classification of one irregular and unaligned pixel time series  $\mathbf{X}^*$  and its associated representation  $\mathbf{Z}$ .

The authors of [25] further propose to use multi-head 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  $\beta_H$  of size  $H$  is used to produce the interpolated value

$$\hat{x}_j(r_l) = \beta_H^\top \Gamma_{r_l}^{H^\top} \mathbf{x}_j^*. \quad (6)$$

with  $\Gamma_{r_l}^H = [\gamma_{r_l}^1, \dots, \gamma_{r_l}^H]$ . 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.

### B. Adaptation of the mTAN for the classification task

Fig. 2 represents the workflow for the classification of one irregular and unaligned pixel time series  $\mathbf{X}^*$  and its associated representation  $\mathbf{Z}$ . In this paper, we propose to use end-to-end learning by combining the mTAN  $h_{\theta_1}$  described in the previous section with the Sparse Variational Gaussian Processes (SVGP) classifier  $f_{\theta_2}$  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  $\theta_1$  and  $\theta_2$  (i.e. the parameters of the mTAN and the SVGP, respectively) and to minimize the error between the predicted class  $\hat{y}$  and the true class  $y$ . This section presents how the mTAN is extended in order to improve the representation  $\mathbf{Z}$  obtained for the classification task.

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. Furthermore, by adding a linear layer after the interpolation, spectral feature reduction can be performed. 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

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

The overall spectro-temporal feature reduction can be written as:

$$\mathbf{Z} = \mathbf{B}\mathbf{X}^*\mathbf{\Gamma} \quad (8)$$

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  $\mathbf{\Gamma} = [\gamma_{r_1}, \dots, \gamma_{r_R}] \in \mathbb{R}^{T \times R}$ .

As defined in (8), the matrix  $\mathbf{\Gamma}$  does not depend on the spectral features and the matrix  $\mathbf{B}$  does not depend on time. Thus, as 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 [26], the spatial coordinates  $(\psi_1, \psi_2)$  are mapped onto a higher dimensional space of dimension  $F$  using  $\varphi$ :

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

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:

$$\tilde{\mathbf{X}}^* = \mathbf{X}^* + \mathbf{P}. \quad (9)$$

The parameters of the perceptron are jointly optimized with the mTAN and the SVGP during the learning step.

The SVGP classifier  $f_{\theta_2}$  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  $\ell$  the lengthscale parameter of the kernel. The square Frobenius norm can be written as

$$\begin{aligned} \|\mathbf{Z}^i - \mathbf{Z}^{i'}\|_F^2 &= \underbrace{\|\mathbf{B}\mathbf{X}^{i*}\mathbf{\Gamma}^i - \mathbf{B}\mathbf{X}^{i'*}\mathbf{\Gamma}^{i'}\|_F^2}_{\text{A}} \\ &\quad + \underbrace{\|\mathbf{B}\mathbf{P}^i\mathbf{\Gamma}^i - \mathbf{B}\mathbf{P}^{i'}\mathbf{\Gamma}^{i'}\|_F^2}_{\text{B}} \\ &\quad + 2 \underbrace{\langle \mathbf{B}(\mathbf{X}^{i*}\mathbf{\Gamma}^i - \mathbf{X}^{i'*}\mathbf{\Gamma}^{i'}), \mathbf{B}(\mathbf{P}^i\mathbf{\Gamma}^i - \mathbf{P}^{i'}\mathbf{\Gamma}^{i'}) \rangle_F}_{\text{C}}. \end{aligned}$$

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

TABLE I: Description of the mTAN parameters  $\theta_1$  and their corresponding sizes. The MLP corresponds to the parameters of a two-layer perceptron used to obtain the spatial positional encoding matrix  $\mathbf{P}$  described in the previous section.

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

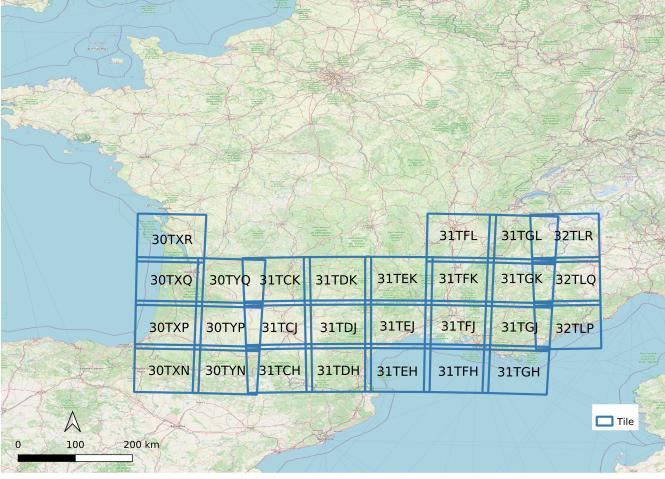


Fig. 3: Location of the 27 studied tiles where a blue square corresponds to one tile as provided by the Theia Data Center<sup>2</sup>. Each tile is displayed with its name in the Sentinel-2 nomenclature (background map © OpenStreetMap contributors).

function  $k(\mathbf{Z}^i, \mathbf{Z}^{i'})$  is composed of an additional element: the interaction term  $\mathbf{C}$ , 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) *Description of the parameters:* The parameters  $\theta_1$  of the mTAN  $h_{\theta_1}$  and their corresponding sizes are summarized in Table I. The total number of learnable parameters  $\theta_1$  is described by the following equation:

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

As a reminder, the parameters  $\theta_2$  of the SVGP classifier  $f_{\theta_2}$  were highly dependent on the number of spectro-temporal features  $T \times D$ . By using an end-to-end training with the mTAN, this number is significantly reduced to  $R \times D'$  with  $R < T$  and  $D' < D$  and therefore the total number of parameters  $\theta_2$  is reduced as well. Numbers of parameters for our data set are given in the experimental section IV-A2.

### III. DATA SET AND EXPERIMENTAL SET-UP

The study area covers approximately 200 000 km<sup>2</sup> in the south of metropolitan France. It is composed of 27 Sentinel-2 tiles, as displayed in Fig. 3.

TABLE II: Number of pixels for each data set

Training	Validation	Test
92 000	23 000	230 000

#### A. Irregular and unaligned Sentinel-2 time series

All available acquisitions of level 2A 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 [29] and were downloaded from the Theia Data Center<sup>2</sup>. All the bands at 20m/pixel were spatially up-sampled to 10m/pixel using bicubic interpolation [30]. A total of 10 spectral bands with three spectral indices (NDVI, NDWI, 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 data sets. The three data sets are class-balanced: 4 000 pixels per class in the *training* data set, 1 000 pixels per class in the *validation* data set and 10 000 pixels per class in the *test* data set. The total number of pixels for each data set is provided in Table II. Classification metrics such as overall accuracy (OA) or F-score were computed for each model using the *test* data set with 9 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* data set and then used to standardize the other data sets (*validation*, *test*) [31].

#### 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 *mTAN-SVGP* model. Two other classifiers  $f_{\theta_2}$  are also compared in terms of classification accuracy and processing time:
























- Multi-layer Perceptron (MLP) with the same setup as in [14],
- Lightweight Temporal Self-Attention (LTAE) described in [13].

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

<sup>2</sup><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

Unlike SVGP or MLP classifiers, the LTAE classifier uses attention mechanisms. It may be redundant to use attention mechanisms both in the mTAN and in the LTAE. Therefore, the LTAE classifier was also studied without the mTAN 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  $\mathbf{m}$  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 *mTAN-SVGP* model. Firstly, 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. Then, the latent representation and the similarity kernel learned by the interpolator are discussed.

##### A. Performance results

1) *Comparison with linear interpolation*: Firstly, the *mTAN-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<sup>3</sup>. Moreover, the number of latent spectral features was equal to the number of spectral features such as  $D' = D = 13$ . Even if there is no reduction, linear mixing is used for the latent spectral features. The latent representation  $\mathbf{Z}$  obtained using the mTAN is described by  $R \times D' = 481$  spectro-temporal features. The *Gapfilled-SVGP*

<sup>3</sup>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.

TABLE IV: Averaged overall accuracies (OA) for the *mTAN-SVGP* and *Gapfilled-SVGP* models (mean %  $\pm$  standard deviation computed with nine runs)

mTAN-SVGP	Gapfilled-SVGP
77.44 $\pm$ 0.15	67.25 $\pm$ 0.37

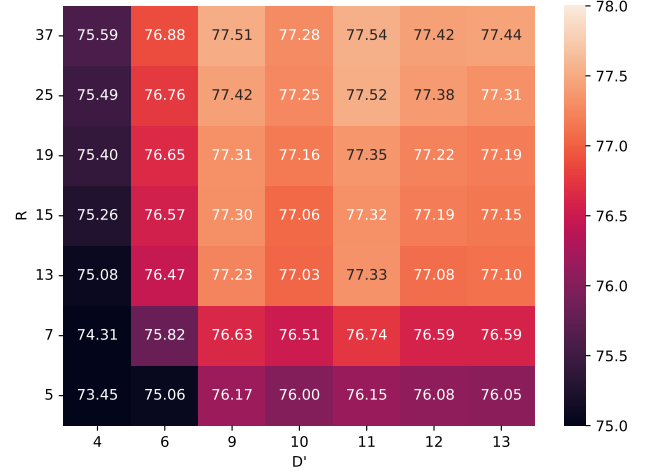


Fig. 4: Averaged overall accuracies (OA) for  $H = 1$  (mean in % computed over 9 different runs) with  $R$  the number of latent dates and  $D'$  the number of latent spectral features.

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 overall accuracy between the *Gapfilled-SVGP* and *mTAN-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 *mTAN-SVGP* model is 10 points above the *Gapfilled-SVGP* model in terms of classification accuracy. The learned latent representation  $\mathbf{Z}$  obtained by the mTAN extracts more meaningful information for the classification task for the SVGP classifier compared to the linearly interpolated data.

2) *Spectral and temporal feature reduction*: Fig. 4 and Fig. 5 represent respectively the averaged overall accuracies (OA) 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  $D' = \{4, 6, 9, 10, 11, 12, 13\}$ . 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). Moreover, the standard deviation of the OA is small, around 0.15. The number of parameters  $\theta_2$  is almost reduced by a factor four (i.e. from 584 200 to 165 600 parameters) and the training times are divided by two, as described in Fig. 5.

In addition, the number of heads  $H$  has a little impact on the classification performances as shown in Fig. 13 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. 14 in Appendix B.

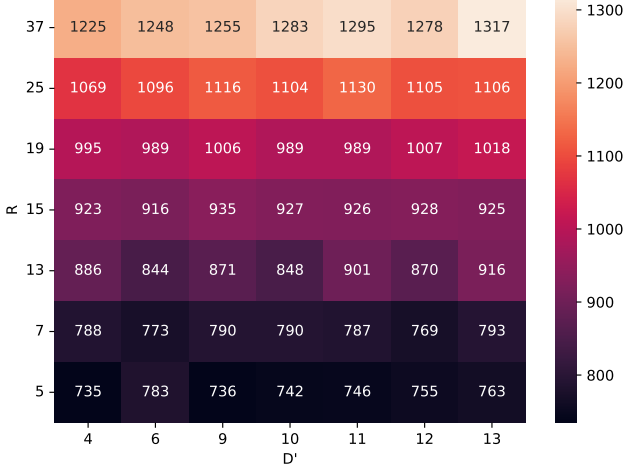


Fig. 5: Averaged training times in seconds for  $H = 1$  (mean computed over 9 different runs) with  $R$  the number of latent dates and  $D'$  the number of latent spectral features.

TABLE V: Averaged overall accuracies (OA) with and without the spatial positional encoded matrix  $\mathbf{P}$  in the mTAN for the model *mTAN-SVGP* (mean %  $\pm$  standard deviation computed with nine runs)

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

3) *Spatial positional encoding*: The spatial information used to compute the positional encoded matrix  $\mathbf{P}$  is composed of the spatial coordinates (northing  $\psi_1$  and easting  $\psi_2$ ) in meters in the Lambert 93 projection. The 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 mTAN for the *mTAN-SVGP* model increased by nearly 1.5 points the overall accuracy. 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. As shown in both Fig. 6a and 6b, the spatial transitions are quite smooth. Each feature takes into account differently the spatial information. Moreover, the learned spatial similarity is anisotropic (see Eq. 9).

4) *Influence on the number of inducing points*: Fig. 7 represents the number of learnable parameters  $\theta_2$  based on the number of spectro-temporal features  $R \times D'$  and the number of inducing points  $M$ . By using spectro-temporal reduction as described in Section IV-A2, the number of spectro-temporal features has been considerably reduced from 481

TABLE VI: Averaged overall accuracies (OA) (mean %  $\pm$  standard deviation) and averaged training times (in sec) for the *mTAN-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).

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

( $R = 37, D' = 13$ ) to 117 ( $R = 13, D' = 9$ ). It results in a significant reduction of the number of learnable parameters  $\theta_2$  as shown in Fig. 7. Moreover, by doubling the number of inducing points from 50 to 100, the number of parameters  $\theta_2$  with  $R = 13, D' = 9$  is still lower than with 50 inducing points and  $R = 37, D' = 13$ .

Experiments were done by increasing the number of inducing points  $M = \{100, 150, 200\}$ . Table VI represents averaged overall accuracies and training times computed with different number of inducing points. With  $M = 200$ , the overall accuracy is almost increased by one point compared to  $M = 50$ . Training time is only slightly affected by this increase in the number of inducing points, i.e. 834s to 967s. 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.

### B. Latent representation

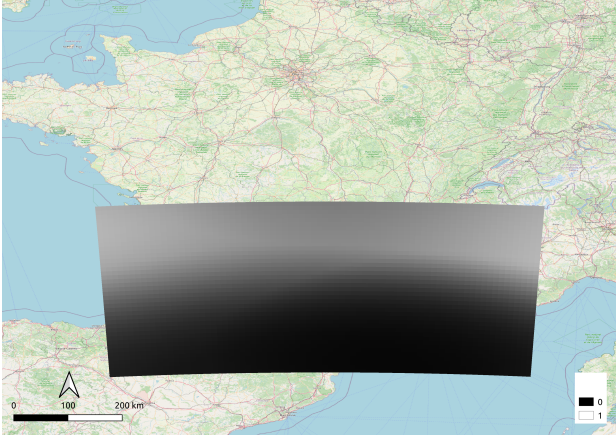
It is possible to visualize the latent representation  $\hat{\mathbf{x}}_j$ . Fig. 8 represents the comparison of three NDVI time series profiles from one pixel labeled as "CORN": the raw data, the gapfilled data (i.e. linearly interpolated) and the latent representation obtained by the mTAN.

The latent mTAN representation obtained in Fig. 8 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.

### C. 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  $\gamma_{r_l}$  defined in (5) for different latent dates  $r_l$  and for different sets of observed dates  $\mathbf{T}$ . In Fig. 9, two different latent dates are studied  $r_l = 181$  and  $r_l = 361^4$ . For each latent date  $r_l$ , two different sets of observed dates  $\mathbf{T}$  are considered. Firstly, the attention value was computed with a regular set of observed dates:  $\mathbf{T} = \{1, \dots, 365\}$  with an interval of  $\tau = 1$  day (in red in Fig. 9). Then, the attention value was computed with a random set of observed dates

<sup>4</sup>The attention value plotted was normalized (cf Fig. 9) in order to have everything on the same vertical scale.



(a) feature 4



(b) feature 12

Fig. 6: Spatial positional encoding  $\mathbf{P}$  computed over a regular grid of spatial coordinates (background map © OpenStreetMap contributors).

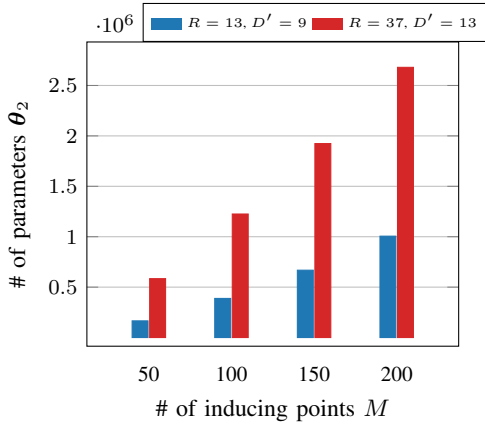


Fig. 7: Number of learnable parameters  $\theta_2$  based on the number of inducing points  $M$  and the number of spectro-temporal features  $R \times D'$ .

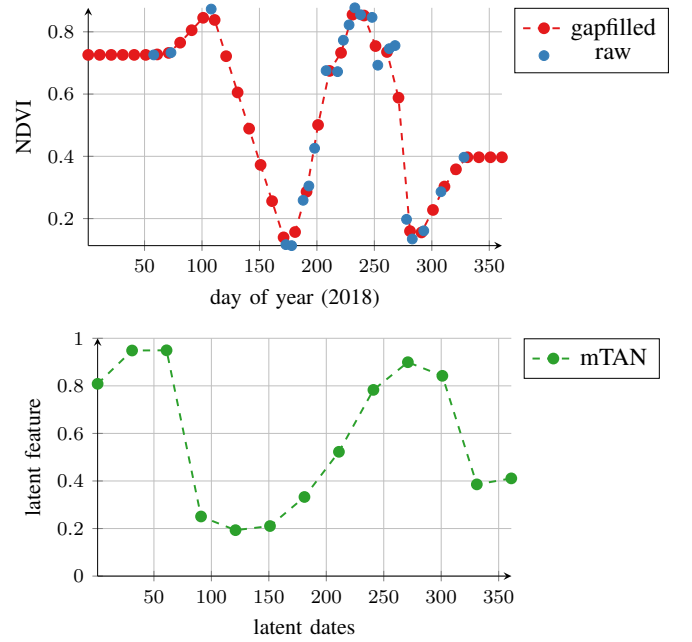


Fig. 8: NDVI time series profiles for a pixel labeled "CORN". Blue points  $\bullet$  correspond to the raw data, the outlier values have been removed in order to have a comprehensive plot. Red points  $\bullet$  correspond to the value obtained with a linear interpolation with an interval of 10 days for a total of 37 dates. Green points  $\bullet$  correspond to the mTAN representation  $\hat{\mathbf{x}}_j$  with  $j = \text{NDVI}$  obtained from the *mTAN-SVGP* model, before the spectral reduction ( $D' = 9$ ).

from a pixel  $i$  with  $\mathbf{T} = \mathbf{T}^i$  (in blue in Fig. 9).

As shown in Fig. 9, the kernel is not centered on the latent date  $r_l$ . It adapts itself according to the latent date  $r_l$  and the available observations. Moreover, as shown in Fig. 9, for the set of observed dates  $\mathbf{T} = \{1, \dots, 365\}$  (i.e. in red), the bandwidth is larger for the latent date  $r_l = 361$  than for the latent date  $r_l = 181$ . Such kernel is referred to as a variable-bandwidth kernel in the statistical literature [32]. It has shown to perform well but, with standard statistical tool, was difficult to optimize. Using the proposed framework, the optimization is efficient.

## V. COMPARISON WITH COMPETITIVE METHODS

This section presents a comparison of the *mTAN-SVGP* model with different models: *mTAN-MLP*, *mTAN-LTAE* and *raw-LTAE*. Firstly, the performance results are studied quantitatively and qualitatively. Then, an additional comparative study is made between the *mTAN-SVGP* and the *raw-LTAE* to evaluate the temporal sampling robustness.

In the following, from the results obtained in the previously in Section IV, the mTAN is set-up with  $R = 13$  latent dates,  $D' = 9$  latent spectral features,  $H = 1$  head, the use of the spatial positional encoding matrix  $\mathbf{P}$  and  $M = 200$  inducing points. As the *raw-LTAE* model is the only one not using mTAN, no spectral or temporal reduction was implemented in this model.

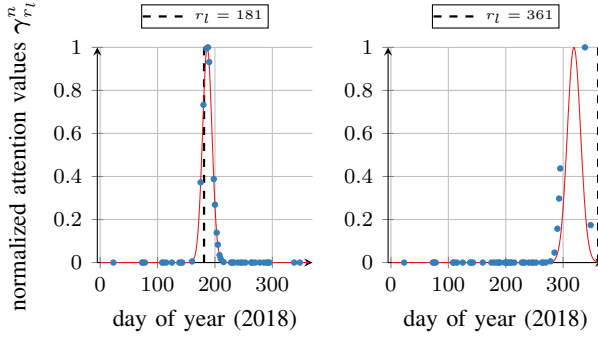


Fig. 9: Normalized attention values  $\gamma_{r_l}^n = \frac{\gamma_{r_l}}{\max(\gamma_{r_l})}$  computed on two different latent dates  $r_l = 181$  and  $r_l = 361$ . — corresponds to  $\gamma_{r_l}^n$  computed with  $\mathbf{T} = \{1, \dots, 365\}$  with a regular interval of  $\tau = 1$  day. Blue points • correspond to  $\gamma_{r_l}^n$  computed with  $\mathbf{T} = \mathbf{T}^i$  for a random pixel  $i$ .

TABLE VII: Averaged training times (in sec) computed over nine runs and number of trainable parameters for each model (*mTAN-SVGP*, *mTAN-MLP*, *mTAN-LTAE* and *raw-LTAE*).

	<i>mTAN-SVGP</i>	<i>mTAN-MLP</i>	<i>mTAN-LTAE</i>	<i>raw-LTAE</i>
Training time	967	1207	840	1279
# parameters	1 005 675	33 113	184 376	761 380

#### A. Performance results

1) *Quantitative results*: Classification accuracies are given in Fig. 10. From the results, the SVGP model took greater advantage of the mTAN than the MLP or the LTAE models. Indeed, the overall accuracy of the *mTAN-SVGP* model is seven points above the *mTAN-MLP* model and around four points above the *mTAN-LTAE* model. On the other hand, the *mTAN-SVGP* model is in average two points below the *raw-LTAE* model. The f-score, recall and precision per class for the *mTAN-SVGP*, *mTAN-MLP*, *mTAN-LTAE* and *raw-LTAE* are represented in Appendix C.

The number of trainable parameters and the training times for each method are summarized in Table VII. The *mTAN-SVGP* model has more trainable parameters than the *raw-LTAE* model. However, the training time of the *mTAN-SVGP* model is about 1.3 times shorter than the *raw-LTAE* as shown in Table VII. By using a spectro-temporal reduction with the mTAN, the number of trainable parameters for the *mTAN-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 the mTAN, for the LTAE, the number of trainable parameters is reduced by four. However, as shown in Fig. 10, the overall accuracy of the *mTAN-LTAE* model is seven points below the *raw-LTAE* model.

2) *Qualitative results*: Land cover maps have been produced for each model (*mTAN-SVGP*, *mTAN-MLP*, *mTAN-LTAE*, *raw-LTAE*) using the *iota*<sup>2</sup> processing chain [33] on two different tiles: 31TCJ and 31TDJ. Inference was performed using the model trained on the 27 tiles with the

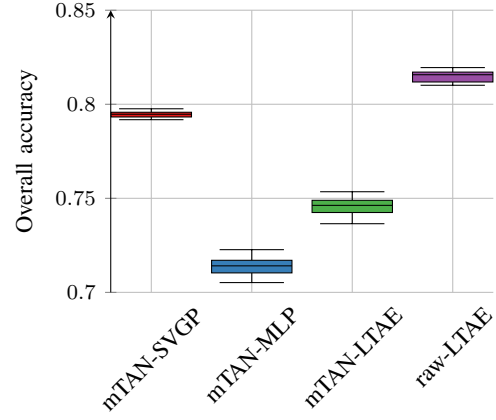


Fig. 10: Boxplots of the overall accuracy for each model (*mTAN-SVGP*, *mTAN-MLP*, *mTAN-LTAE* and *raw-LTAE*) computed over nine runs.

best overall accuracy over the nine runs. The results obtained by respectively the *mTAN-SVGP* and *raw-LTAE* are shown in Fig. 11. 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 [34]. The *mTAN-SVGP* takes into account spatial information without spatial oversmoothing.

All the generated land cover maps are available for download.<sup>5</sup>

#### B. Robustness to the temporal sampling

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 the common temporal grid used during the training step (i.e. the whole set of observed dates  $\mathbf{T} = \{t_1, \dots, t_T\}$ ). It is not the case for classifiers using the mTAN. Thus, once trained, they are able to classify any irregular and unaligned pixel time series.

To study the robustness to the temporal sampling, dates not seen during the training step were artificially created. They correspond to the acquisition dates  $\mathbf{T}$  for the *test* data set that have been slightly shifted. 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 overall accuracy was computed only on 31TCJ tile for two models *mTAN-SVGP* and *raw-LTAE* both trained on the 27 tiles. As shown in Fig. 12, the overall accuracy of the *mTAN-SVGP* model is not affected by this temporal shift  $\delta$ . However, the OA of the *raw-LTAE* model is greatly impacted by the temporal shift  $\delta$  and the OA is almost divided by 1.5 with  $\delta = 5$  days. By using a linear smoother with temporal attention mechanisms, the *mTAN-SVGP* model is more robust to this shift than the *raw-LTAE* model which use spectro-temporal attention mechanisms. Thus, the *raw-LTAE* is more sensitive

<sup>5</sup>DOI: <https://doi.org/10.5281/zenodo.8033902>



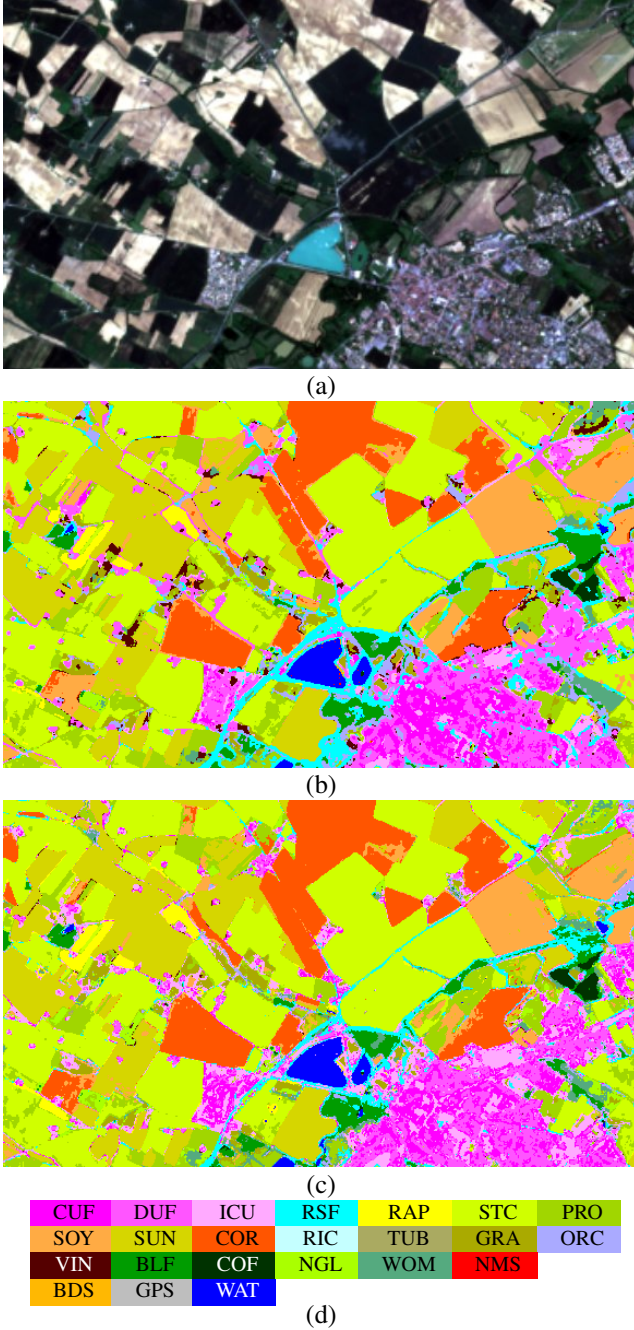


Fig. 11: Land covers maps. (a) Sentinel-2 true color composition, (b) Classification map obtained with *mTAN-SVGP*, (c) Classification map obtained with *raw-LTAE* and (d) the class color map (see Table III for correspondence).

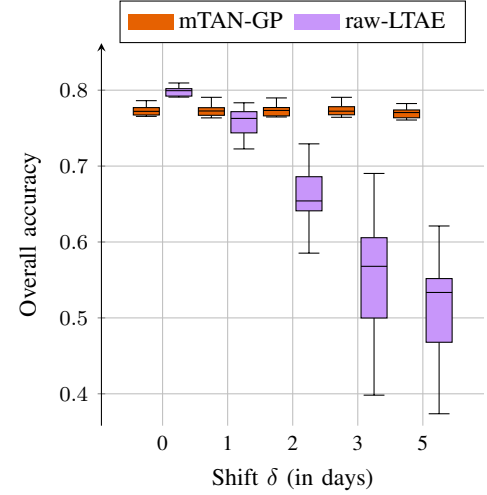


Fig. 12: Boxplots of the overall accuracy for the *mTAN-SVGP* and *raw-LTAE* models computed with the *test* data set only on the 31TCJ tile over nine runs. The models were trained and validated on the all 27 tiles. The acquisition dates  $T$  for the test data set were artificially shifted with different values:  $\delta = \{0, 1, 2, 3, 5\}$  days.

to the presence of dates during the training step and may therefore be likely to over-fit.

## VI. CONCLUSIONS AND PERSPECTIVES

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 interpolate and learn strategy is proposed. A first module, the Multi-Attention Time Networks (*mTAN*), enables to project the irregular and unaligned SITS onto a fixed and reduced size representation. This representation is then given to the SVGP classifier and all the parameters are jointly optimized during the optimization of the classifier. The spatial information is taken into account in the representation through the *spatial positional encoding*. Experiments were conducted on Sentinel-2 SITS of the full year 2018 in the south of France.

In terms of accuracy, the end-to-end learning *mTAN-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 *mTAN-SVGP* model is above the *mTAN-MLP* and *mTAN-LTAE* models in terms of accuracy. While the proposed *mTAN-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 very sensitive to variations in the set of available dates during inference, contrary to the proposed *mTAN-SVGP* which showed stable performances. The *mTAN-SVGP* is therefore more likely to generalize well to large scale scenario where irregular and variable sampling dates are prominent.

In this paper, the potential of the multi-head 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  $\beta_H$  in Eq. (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 multi-modal 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 mTAN to process unaligned time series would make the fusion of multi-sensor 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 (*mTAN-GP*, *mTAN-MLP*, *mTAN-LTAE* and *raw-LTAE*) is made available in the following repository: [https://gitlab.cesbio.omp.eu/belletv/land\\_cover\\_southfrance\\_mt看\\_gp\\_irregular\\_sits](https://gitlab.cesbio.omp.eu/belletv/land_cover_southfrance_mt看_gp_irregular_sits).

#### ACKNOWLEDGMENT

The authors would like to thank Benjamin Tardy, from CS Group - France, for his support and help during the generation of the different data sets and the production of land cover classification maps with the *iota*<sup>2</sup> software. Finally, the authors would also like to thank CNES for the provision of its high performance computing (HPC) infrastructure to run the experiments presented in this paper and the associated help.

#### REFERENCES

- [1] IPCC, “SYNTHESIS REPORT OF THE IPCC SIXTH ASSESSMENT REPORT (AR6) longer report.” [https://report.ipcc.ch/ar6syr/pdf/IPCC\\_AR6\\_SYR\\_LongerReport.pdf](https://report.ipcc.ch/ar6syr/pdf/IPCC_AR6_SYR_LongerReport.pdf), 2023.
- [2] C. Persello, J. D. Wegner, R. Hansch, D. Tuia, P. Ghamisi, M. Koeva, and G. Camps-Valls, “Deep Learning and Earth Observation to Support the Sustainable Development Goals: Current Approaches, Open Challenges, and Future Opportunities,” *IEEE Geoscience and Remote Sensing Magazine*, pp. 2–30, 2022.
- [3] D. Tuia, K. Schindler, B. Demir, G. Camps-Valls, X. X. Zhu, M. Kochupillai, S. Džeroski, J. N. van Rijn, H. H. Hoos, F. D. Frate, M. Datcu, J.-A. Quiané-Ruiz, V. Markl, B. L. Saux, and R. Schneider, “Artificial intelligence to advance earth observation: a perspective,” 2023.
- [4] F. Bertini, O. Brand, S. Carlier, U. Del Bello, M. Drusch, R. Duca, V. Fernandez, C. Ferrario, M. H. Ferreira, C. Isola, V. Kirschner, P. Laberinti, M. Lambert, G. Mandorlo, P. Marcos, P. Martimort, S. Moon, P. Oldeman, M. Palomba, and J. Pineiro, “Sentinel-2 esa’s optical high-resolution mission for gmes operational services,” *ESA bulletin. Bulletin ASE. European Space Agency*, vol. SP-1322, 03 2012.
- [5] G. Camps-Valls, “Machine learning in remote sensing data processing,” in *2009 IEEE International Workshop on Machine Learning for Signal Processing*, 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 Processing Magazine*, vol. 31, no. 1, pp. 45–54, 2014.
- [7] X. X. Zhu, D. Tuia, L. Mou, G.-S. Xia, L. Zhang, F. Xu, and F. Fraundorfer, “Deep learning in remote sensing: A comprehensive review and list of resources,” *IEEE Geoscience and Remote Sensing Magazine*, vol. 5, no. 4, pp. 8–36, 2017.
- [8] M. Belgiu and L. Drăguț, “Random forest in remote sensing: A review of applications and future directions,” *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 114, pp. 24–31, Apr. 2016.
- [9] J. Inglada, A. Vincent, M. Arias, B. Tardy, D. Morin, and I. Rodas, “Operational High Resolution Land Cover Map Production at the Country Scale Using Satellite Image Time Series,” *Remote Sensing*, vol. 9, no. 1, 2017.
- [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 Sensing of Environment*, vol. 187, pp. 156–168, 2016.
- [11] M. Rußwurm and M. Körner, “Multi-temporal land cover classification with sequential recurrent encoders,” *ISPRS International Journal of Geo-Information*, vol. 7, no. 4, 2018.
- [12] C. Pelletier, G. Webb, and F. Petitjean, “Temporal Convolutional Neural Network for the Classification of Satellite Image Time Series,” *Remote Sensing*, vol. 11, p. 523, Mar. 2019.
- [13] V. Sainte Fare Garnot and L. Landrieu, “Lightweight temporal self-Attention for classifying satellite images time series,” in *Workshop on Advanced Analytics and Learning on Temporal Data, AALTD*, Sept. 2020.
- [14] V. Bellet, M. Fauvel, and J. Inglada, “Land cover classification with gaussian processes using spatio-spectro-temporal features,” *IEEE Transactions on Geoscience and Remote Sensing*, pp. 1–1, 2023.
- [15] S. C.-X. Li and B. Marlin, “A scalable end-to-end gaussian process adapter for irregularly sampled time series classification,” in *Proceedings of the 30th International Conference on Neural Information Processing Systems, NIPS’16*, p. 1812–1820, Curran Associates Inc., 2016.
- [16] A. Constantin, M. Fauvel, and S. Girard, “Joint Supervised Classification and Reconstruction of Irregularly Sampled Satellite Image Times Series,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, p. 4403913, May 2021.
- [17] F. Petitjean, J. Inglada, and P. Gancarski, “Satellite image time series analysis under time warping,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 50, no. 8, pp. 3081–3095, 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,” *Journal of Statistical Software*, vol. 88, no. 5, p. 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 Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 9, no. 8, pp. 3729–3739, 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," *Journal of Information and Data Management*, vol. 10, p. 16–34, Jun. 2019.
- [21] D. Neil, M. Pfeiffer, and S.-C. Liu, "Phased Istm: Accelerating recurrent network training for long or event-based sequences," 2016.
- [22] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems* (I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, eds.), vol. 30, Curran Associates, Inc., 2017.
- [23] M. Rußwurm and M. Körner, "Self-attention for raw optical satellite time series classification," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 169, pp. 421–435, 2020.
- [24] S. N. Shukla and B. M. Marlin, "Interpolation-prediction networks for irregularly sampled time series," in *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*, OpenReview.net, 2019.
- [25] S. N. Shukla and B. Marlin, "Multi-time attention networks for irregularly sampled time series," in *International Conference on Learning Representations*, 2021.
- [26] L. Baudoux, J. Inglada, and C. Mallet, "Toward a Yearly Country-Scale CORINE Land-Cover Map without Using Images: A Map Translation Approach," *Remote Sensing*, vol. 13, p. 1060, Mar. 2021.
- [27] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*. Springer Series in Statistics, New York, NY, USA: Springer New York Inc., 2001.
- [28] A. Zhang, Z. C. Lipton, M. Li, and A. J. Smola, "Dive into deep learning," *arXiv preprint arXiv:2106.11342*, 2021.
- [29] 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 Sensing*, vol. 11, p. 433, Feb. 2019.
- [30] O. D. Team, "Orfeo ToolBox 7.1," Mar. 2020. <https://zenodo.org/record/3715021>.
- [31] M. Kuhn and K. Johnson, *Feature Engineering and Selection: A Practical Approach for Predictive Models*. Chapman and Hall/CRC, 1 ed., July 2019.
- [32] G. R. Terrell and D. W. Scott, "Variable Kernel Density Estimation," *The Annals of Statistics*, vol. 20, no. 3, pp. 1236 – 1265, 1992.
- [33] J. Inglada, A. Vincent, M. Arias, and B. Tardy, *iota2-a25386*, July 2016. <https://doi.org/10.5281/zenodo.58150>.
- [34] 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 Sensing*, vol. 11, p. 1986, Aug 2019.



**Valentine Bellet** (Graduate Student Member, IEEE) received the master's degree in automatic control and electronics from INSA Toulouse, Toulouse, France, in 2019. She is currently pursuing the Ph.D. degree with the Centre d'Études Spatiales de la Biosphère (CESBIO) Laboratory, Université de Toulouse, Toulouse. She is working on the subject of land cover pixel-based classification with satellite image time series (SITS) at a national scale.



**Mathieu Fauvel** 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 at INRAe and Centre d'Études Spatiales de la Biosphère (CESBIO) Laboratory, Université de Toulouse, Toulouse. His research interests are remote sensing, machine learning, and image processing.



**Jordi Inglada** 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** Julien Michel received the Telecommunications Engineer degree from the École Nationale Supérieure des Télécommunications de Bretagne, Brest, France, in 2006. He is currently with the Centre National d'Études Spatiales (French Space Agency), Université de Toulouse, Toulouse, France, where he is working as a research engineer on remote sensing image processing at the Centre d'Études Spatiales de la Biosphère (CESBIO). He is also currently pursuing the Ph.D. degree with the Centre d'Études Spatiales de la Biosphère (CESBIO), Université de Toulouse, Toulouse. 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.

## APPENDIX A

### SOLVER PARAMETERS FOR EACH MODEL

TABLE VIII: Parameter values for the Adam optimizer for the models: *Gapfilled-SVGP*, *mTAN-SVGP*, *mTAN-MLP*, *mTAN-LTAE* and *raw-LTAE*.

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

## APPENDIX B

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



Fig. 13: Averaged overall accuracies (OA) (mean in % computed over 9 different runs) with  $R$  the number of latent dates,  $D'$  the number of latent spectral features and  $H$  the number of heads.

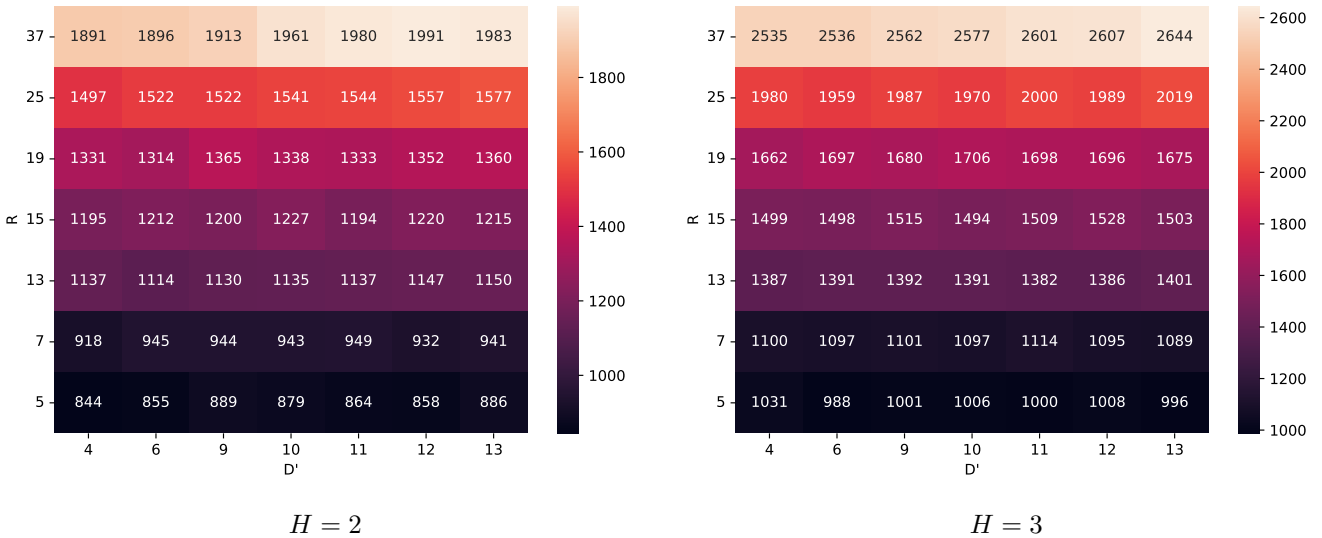


Fig. 14: Averaged training times in seconds (mean computed over 9 different runs) with  $R$  the number of latent dates,  $D'$  the number of latent spectral features and  $H$  the number of heads.

### APPENDIX C METRICS PER CLASS

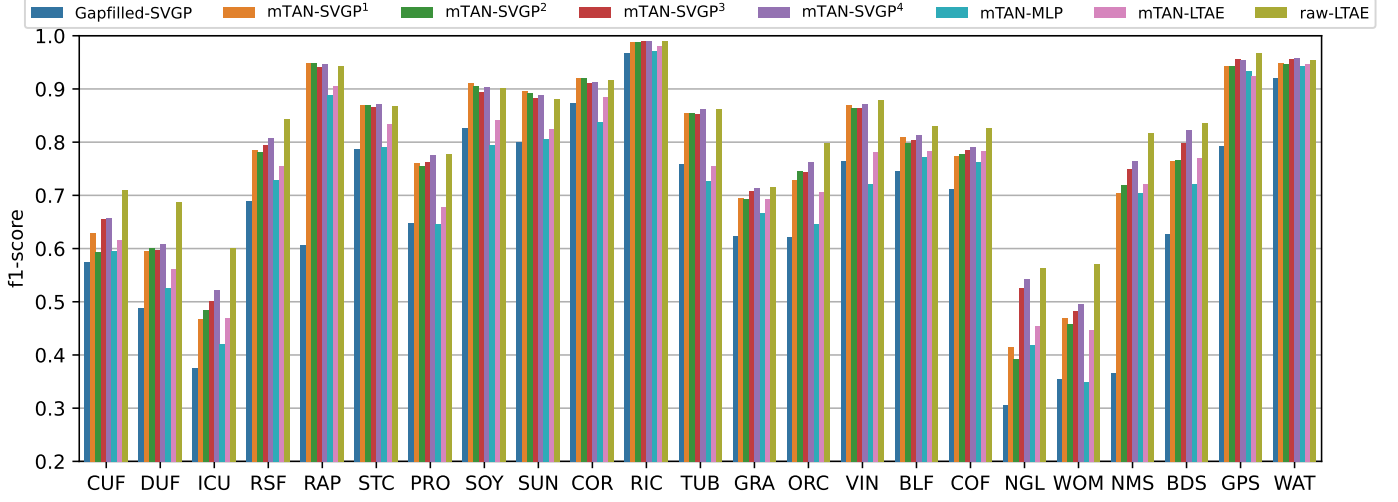


Fig. 15: Averaged f-score per class. mTAN-SVGP<sup>1</sup>:  $H = 1, D' = 13, R = 37, M = 50$  ; mTAN-SVGP<sup>2</sup>:  $H = 1, D' = 9, R = 13, M = 50$  ; mTAN-SVGP<sup>3</sup>: with **P** and  $H = 1, D' = 9, R = 13, M = 50$ ; mTAN-SVGP<sup>4</sup>: with **P** and  $H = 1, D' = 9, R = 13, M = 200$ .

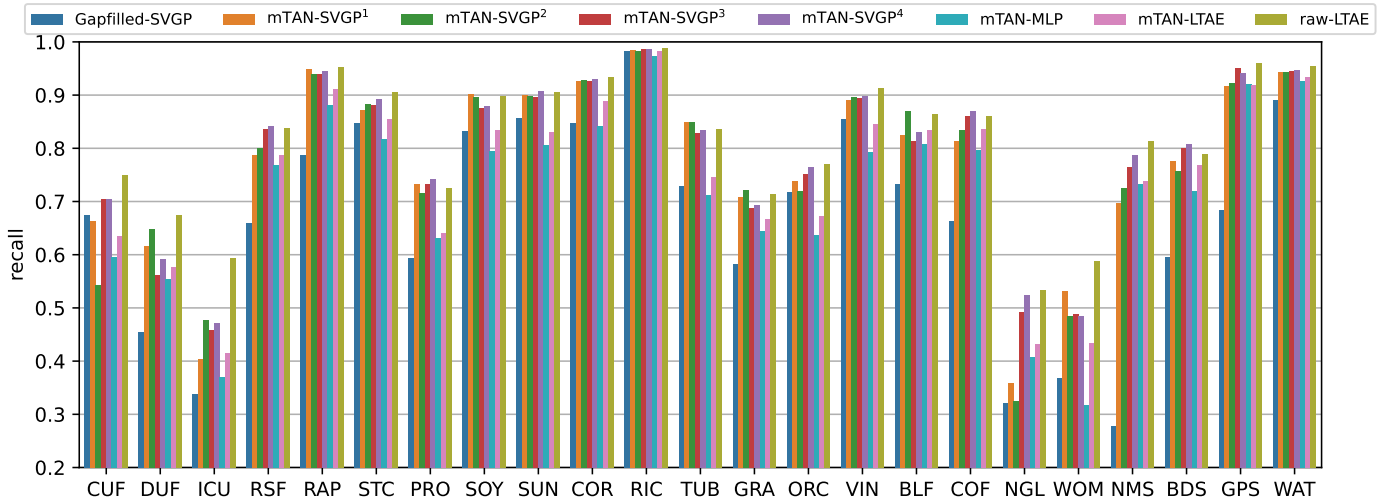


Fig. 16: Averaged recall per class. mTAN-SVGP<sup>1</sup>:  $H = 1, D' = 13, R = 37, M = 50$  ; mTAN-SVGP<sup>2</sup>:  $H = 1, D' = 9, R = 13, M = 50$  ; mTAN-SVGP<sup>3</sup>: with **P** and  $H = 1, D' = 9, R = 13, M = 50$ ; mTAN-SVGP<sup>4</sup>: with **P** and  $H = 1, D' = 9, R = 13, M = 200$ .

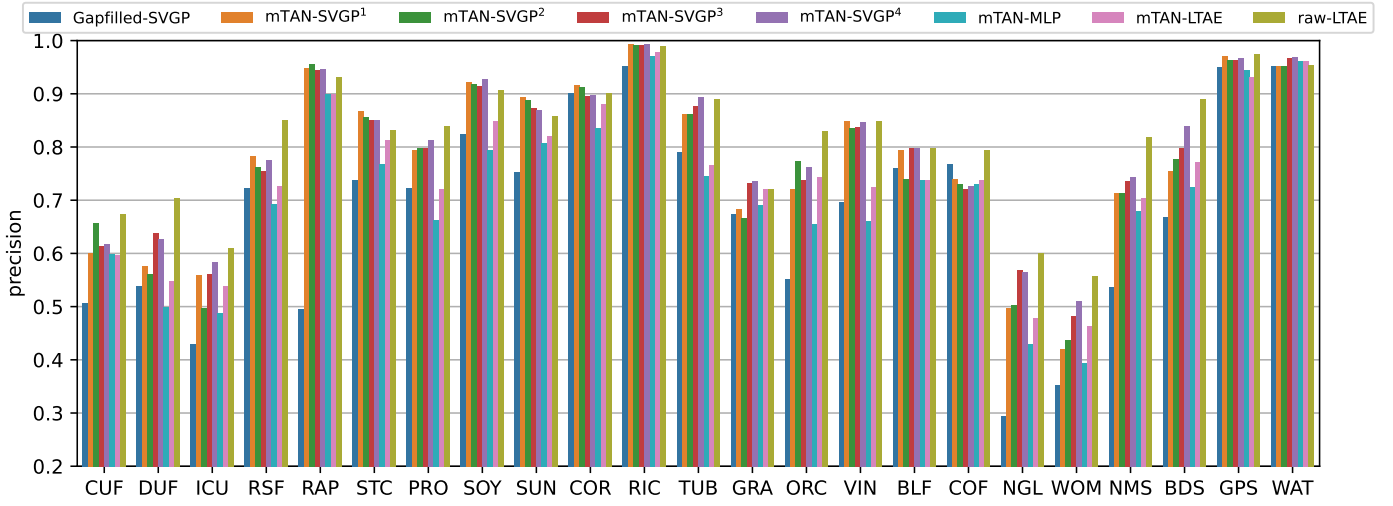


Fig. 17: Averaged precision per class. mTAN-SVGP¹:  $H = 1, D' = 13, R = 37, M = 50$  ; mTAN-SVGP²:  $H = 1, D' = 9, R = 13, M = 50$  ; mTAN-SVGP³: with  $\mathbf{P}$  and  $H = 1, D' = 9, R = 13, M = 50$ ; mTAN-SVGP⁴: with  $\mathbf{P}$  and  $H = 1, D' = 9, R = 13, M = 200$ .