BreizhCrops: A Satellite Time Series Dataset for Crop Type Identification

Marc Rußwurm [*1] **Sébastien Lefèvre** [2] **Marco Körner** [1]

Abstract

This dataset challenges the time series community with the task of satellite-based vegetation identification on large scale real-world dataset of satellite data acquired during one entire year. It consists of time series data with associated crop types from 580k field parcels in Brittany, France (*Breizh* in local language). Along with this dataset, we provide results and code of a Long Short-Term Memory network and Transformer network as baselines. We release dataset, along with preprocessing scripts and baseline models in https://github.com/TUM-LMF/BreizhCrops and encourage methodical researchers to benchmark and develop novel methods applied to satellite-based crop monitoring.

1. Earth Observation and Agricultural Monitoring

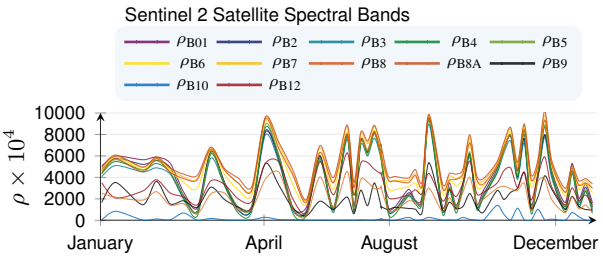
Today, optical satellites observe the entire surface Earth at weekly intervals and measure the reflectance of sunlight at multiple spectral wavelengths. This regular coverage allows for the monitoring of vegetation at discrete time intervals at spatial resolutions between 10m and 60m. The unit of measurement on optical satellite imagery is the surface reflectance

$$\rho_\lambda = \frac{\pi L_\lambda d^2}{E_{\text{sun}} \cos(\varphi_{\text{sun}})} \quad (1)$$

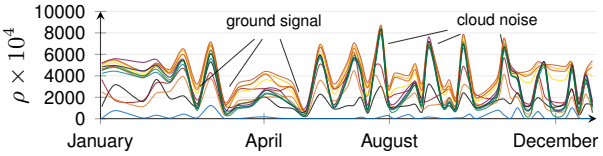
of several distinct wavelength bands indicated by the central wavelength λ and discretized in a spatial grid of pixels. The reflectance is obtained from the measured radiance L_λ in $\frac{W}{srm^2}$ on the satellite sensor projected on the surface using the solar zenith angle φ_{sun} of each pixel. To obtain unitless reflectances, it is normalized with the total solar irradiance E_{sun} in $\frac{W}{srm^2}$ scaled by the squared Earth-Sun distance d (Richter et al., 2010).

[1]Technical University of Munich [2]Université de Bretagne Sud. Correspondence to: Marc Rußwurm <marc.russwurm@tum.de>.

Accepted to the Time Series Workshop of the 36^{th} *International Conference on Machine Learning*, Long Beach, California, PMLR 97, 2019.



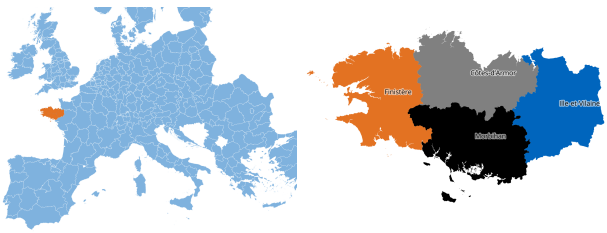
(a) A single example of category *temporary meadows*



(b) A single example of category *corn*

Figure 1: Examples of the input time series of reflectances ρ for all 13 spectral bands of the Sentinel 2 satellite.

The identification of crop types form spaceborne imagery forms an important component of agricultural monitoring. Assessing the cultivated crop types early in the season allows for estimating the expected crop yield at large scale. A classification model trained on crop type identification likely indirectly learns a model of the vegetation phenology, *i.e.*, characteristic life cycle events. Analyses of these events at different regional or temporal scales may in future help estimate the effect of anomalous events, such as droughts, pests on the expected crop yield.



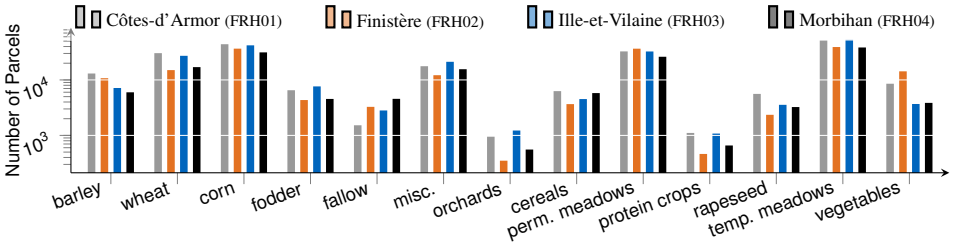
(a) Brittany within Europe (b) NUTS-3 regions of Britanny

Figure 2: The location of the NUTS-3 region *FRH0* of Brittany, France.

.

Departements	NUTS-3	Parcels
Morbihan	FRH04	158522
Côtes-d'Armor	FRH01	221095
Finistère	FRH02	180565
Ille-et-Vilaine	FRH03	207993

(a) The four NUTS-3 Departements of Brittany and number of field parcels per region



(b) Class frequencies per partition

Figure 3: Analyses of the number of parcels and class frequencies per partition in the vector dataset.

2. The Breizh Crops Dataset

The dataset comprises 580k field parcels in the Region of Brittany (Breizh in the local language), France, of the season 2017. We show the region along with two examples in Fig. 2. The example field parcels in Fig. 1b and in Fig. 1a show all spectral information within the respective field geometry within the season of 2017. Note that the data is positively biased by clouds which cause systematically positive outliers in the time series data, as annotated in Fig. 1b

2.1. Organization

The data is organized at a regional level by the *Nomenclature des unités territoriales statistiques (NUTS)* which forms an international standard for referencing authoritative districts. Brittany is a NUTS-2 region, as highlighted in Fig. 2a. It is further divided into the four NUTS-3 regions *Côtes-d'Armor*, *Finistère*, *Ille-et-Vilaine*, and *Morbihan*, shown in Fig. 2b. We partitioned all acquired field parcels according to the NUTS-3 regions and suggest partitioning the dataset in partitions for training, validation, and evaluation based on these spatially distinct regions.

2.2. Satellite Data

To obtain the satellite data, we downloaded *all* available *Sentinel 2* images from *Google Earth Engine* (Gorelick et al., 2017) at processing level L1C. All $D = 13$ spectral bands located within one field parcel were mean-aggregated to a feature vector $\boldsymbol{x}_t$, as shown in the examples in Fig. 2. We provide a script for downloading the data in the associated codebase.

2.3. Crop Type Labels

The *Common Agricultural Policy* of the European Union subsidizes farmers based on the cultivated crops. Each member country is required to gather geographical information of geometry and cultivated crop. This information is obtained from the farmers themselves by surveys within the subsidy application process. National agencies monitor the correctness either by gathering control samples on-site or by means of remote sensing and Earth observation. In France, the *National Institue of Forest and Geography Information (IGN)* is responsible for gathering this information and recently started releasing anonymized parcel geometries and type of cultivated crop with an *open license* policy[1].

For this dataset, we selected the 13 crop groups that appear at least 250 times in each NUTS-3 region and at least 1000 times in all regions. Still, due to the nature of agricultural production focused on a few dominant crop types, a class imbalance can be observed in the data. Regional differences in environmental conditions further vary the label distributions for the respective partitions, as can be seen in the histogram of classes per region in Fig. 3b.

3. Baseline Methods

We implemented two baseline methods to obtain an empiric evaluation on expected classification accuracies on the datasets.

3.1. Multilayer LSTM

First, we employ a multi-layer bidirectional *Long Short-Term Memory (LSTM)* (Hochreiter & Schmidhuber, 1997) encoder that iteratively extracts classification-characteristic features from a sequence of inputs. Throughout this work, we use three stacked bidirectional long short-term memory layers with 128 hidden dimensions.

3.2. Transformer Encoder

Inspired by the success of self-attention networks, we also show classification results using an encoder of the *Transformer* network (Vaswani et al., 2017). It uses N stacked modules of H multiple self-attention heads and d_{model} hidden states within the self attention vectors. We initially experimented with the configuration of the original implementation, *i.e.*, $H = 8, N = 6, d_{model} = 512$, but found it

[1] https://www.data.gouv.fr/en/datasets/registre-parcellaire-graphique-rpg-contours-des-parcelles-et-ilots-culturaux-et-leur-groupe-de-cultures-majoritaire/

difficult to converge to a good solution. Hence, we opted for a smaller network configuration with $N = 4$ layers, $H = 4$ heads, and $d_{model} = 128$.

3.3. Training Details

We trained the network hyper-parameters on the FRH01 and FRH02 partitions of the dataset and observed the accuracy on the FRH03 partition. Each training sequence has a different sequence length. The observation period, however, ranges over one entire year. Hence, we decided to randomly sample $T' = 45$ observations from all available observations while maintaining the sequence order. This sampling strategy results in sequence lengths of fixed length which simplified the batching process and introduces a certain variability which may avoid overfitting of the model on single specific sequence elements. Also, we use the Adam (Kingma & Ba, 2014) optimizer with the learning rate scheduler of (Vaswani et al., 2017). It initializes the learning rate with $\sqrt{d_{model}}$ and increases it linearly for w *warm-up* epochs. After the warm-up phase, the learning rate decreases exponentially. All baselines were trained using cross-entropy loss between the one-hot representation of the predicted crop label and the ground truth.

4. Baseline Results

For the final accuracy evaluation we trained the baselines on the FRH01, FRH02, and FRH03 partitions and report results on the FRH04 region. In Table 1, we compare the two baselines using the overall accuracy, the kappa correlation metric κ (Cohen, 1960), and the class-mean recall, precision, and f_1-score. From these comparisons, both baselines achieved comparable accuracies. The LSTM model slightly outperformed the transformer baseline on the class-averaged metrics, while the transformer was slightly superior in the sample-wise accuracy. Still, the differences between these two baselines were marginal. We analyze the class-wise accuracies further in Fig. 4 to get a detailed insight into the nature of classification performances achieved in this experiment. In Fig. 4a we show the precision, recall, an f1-score for each category in along with the number of samples per category. The mean metrics and sum of parcels are displayed in Fig. 4. Some categories, such as *barley*, *wheat*, *corn*, or *rapeseed* were well-classified with accuracies around 90% while others, such as *protein crops*, *orchards*, or *temporary meadows* were poorly classified. Classes that were composed of single distinct types of vegetation seemed more distinguishable, while broadly defined categories, such as *orchards*, were classified with less accuracy. Note that these categories are official crop culture groups of the French parcel system. Hence, we specifically did not remove these broad categories from the dataset, as they pose a challenging task for methods and match the official categorization system. Beyond the distinction of broad and narrow categories, some common categories, such as *wheat* (2), or *corn* (3) were better classified, while less-common ones, such as *orchards* (7), or *protein crops* (10) were predicted less accurately. We interpret this as behavior caused the imbalance of the dataset. When we observe the confusion matrix in column-normalized recall in Fig. 4c and in row-normalized precision in Fig. 4b, we can observe characteristic confusions between the individual categories.

The accuracy metrics Fig. 4a are shown in in the diagonal, while off-diagonal elements indicate systematic confusions between two classes. Here, it appears that the categories *fodder* (4), *gel* (5), *miscellaneous* (6), and *orchards* (7) were often misclassified with each other. Overall, these results show the complexity of the task of crop-type mapping from remote sensing imagery. While some narrow categories are classified well by the baselines other broader categories achieve poor accuracy metrics.

5. Challenges

The results of the previous section show the general feasibility but demonstrate the complexity of the task of classifying crop types from satellite imagery. This large scale real-world dataset has the potential to impact vegetation monitoring at a European or global scale. The satellite data is globally available, while the ground truth labels are gathering within the European Union. In the following, we outline a series of methodical challenges associated with the dataset that pose demanding questions to the time series community and likely need to be addressed to improve the accuracy of methods trained on this dataset.

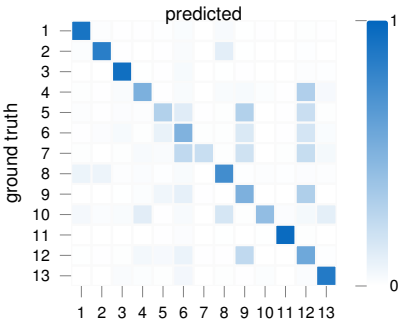
Imbalanced class labels. Agricultural areas are commonly dominated by few common crops, such as *corn*, *meadow*, or *wheat* which are cultivated extensively. Nevertheless, other types of vegetation are still of interest for the local

Table 1: Accuracy metrics for the Multi-layer bidirectional LSTM (Hochreiter & Schmidhuber, 1997) and the Transformer-Encoder (Vaswani et al., 2017).

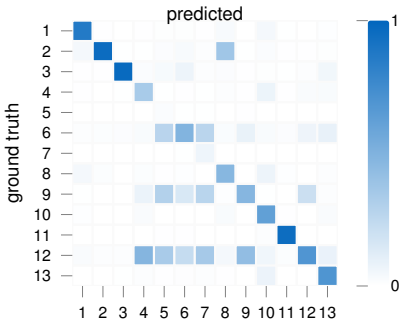
baseline	accuracy	κ	mean f1	mean precision	mean recall
Transformer (Vaswani et al., 2017)	**0.69**	**0.63**	0.57	0.60	0.56
LSTM (Hochreiter & Schmidhuber, 1997)	0.68	0.62	**0.59**	**0.63**	**0.58**

#	crop type	prec.	rec.	f_1	#samples
1	barley	90	86	88	4982
2	wheat	83	95	89	13850
3	corn	93	**96**	94	25059
4	fodder	51	34	41	3449
5	fallow	30	2	4	3863
6	misc.	50	49	49	12499
7	orchards	21	7	10	391
8	cereals	74	47	57	4645
9	perm. meadows	51	47	49	20966
10	protein crops	42	61	50	498
11	rapeseed	**96**	94	**95**	2664
12	temp. meadows	56	68	62	29977
13	vegetables	86	69	76	3114
		63	**58**	**59**	125957

(a) per-class accuracy metrics



(b) precision



(c) recall

Figure 4: Accuracies and class confusions of the bidirectional LSTM-RNN model

authorities and should be classified at a reasonable accuracy. This introduces a strong imbalance in the class frequencies, as shown in Fig. 3b. Please note the logarithmic scale.

Classes with similar characteristics. Some categories can be traced to one unique type of crop, such as *wheat*, or *corn*. Here, the phenological characteristics can be traced to single specific crop types. Other, less frequent classes, are aggregated into groups that incorporate a broader range of vegetation types which may be difficult to distinguish, such as *orchards*.

Non-Gaussian noise induced by clouds. Clouds cover the Earth's surface at regular intervals. Their large reflectance introduces a positive non-gaussian noise to the data at single intervals. This manifests itself by positive outliers in the reflectance data over the time scale, as can be seen in the examples of Fig. 2.

Regional variations in the class distributions. Regional variances in soil quality, elevation, temperature, and precipitation lead to a spatial correlation in the frequency of dominated agricultural crop. This effect increases at larger scales where these environmental conditions change significantly. Still, certain variations in crop distributions based on regionally distinct regions can be seen in Fig. 3b.

Irregular sampling distance. The Sentinel 2 constellation consists of two satellites which orbit the Earth at opposite orbits. In optimal circumstances, one point on the Earth is observed every two to five days. However, due to errors in the data acquisition or bottlenecks in the data downlink, single observations can be skipped. This leads to irregular time intervals between observations in the time series between two and ten days.

Variable sequence length. Earth observation satellites scan the surface in stripes of 290km width (termed *swath*). To ensure a constant coverage, the acquisition is planned with a certain degree overlap towards the border of these stripes. Due to this configuration, the sequence lengths T of acquired images per field parcels are approximately 50 or 100 observations.

Spatial autocorrelation. Spatially closer objects are more similar than distant ones (Tobler, 1970). This autocorrelation can introduce a dependence between training and validation datasets that may disguise overfitting and impede generalization. To counteract this, several researchers (Rußwurm & Körner, 2017; Jean et al., 2018) have adopted a training/validation/evaluation partitioning that groups spatially distant parcels. Hence, we organized the data in their respective NUTS-3 regions to encourage training on these spatially separate regions.

6. Summary and Outlook

In this work, we challenge the time series community with a large scale time series dataset for crop-type mapping. The dataset is gathered, organized and structured from open source raw label data and satellite measurements. The objective is to accurately classify a set of narrow and broad crop type categories in multiple regions in Brittany, France.

A model that internalizes the discriminative characteristics of these crop types encodes the effect of specific life-cycle events of the crop. Hence, methods that excel at this dataset may be good candidates for extracting other vegetation-related characteristics, such as crop yield, or may be suitable to predict droughts. The satellite data of this kind is globally available free of charge which may allow a global application of well-performing model variants in the future.

Beyond the application-wise advantages of this dataset, it also poses a series of methodical challenges to the time series community, as summarized in Section 5. The raw vector dataset and the time series data is available via `https://github.com/tum-lmf/BreizhCrops` along with scripts for preprocessing, and the baseline implementations.

References

Cohen, J. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46, 1960.

Gorelick, N., Hancher, M., Dixon, M., Ilyushchenko, S., Thau, D., and Moore, R. Google earth engine: Planetary-scale geospatial analysis for everyone. *Remote Sensing of Environment*, 202: 18–27, 2017.

Hochreiter, S. and Schmidhuber, J. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

Jean, N., Wang, S., Azzari, G., Lobell, D., and Ermon, S. Tile2vec: Unsupervised representation learning for remote sensing data. *arXiv preprint arXiv:1805.02855*, 2018.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Richter, R., Louis, J., and Müller-Wilm, U. Sentinel-2 MSI—Level 2A Products Algorithm Theoretical Basis Document. Technical report, Deutsches Zentrum für Luft- und Raumfahrt e.V. (DLR), VEGA Technologies SAS, 2010.

Rußwurm, M. and Körner, M. Temporal vegetation modelling using long short-term memory networks for crop identification from medium-resolution multi-spectral satellite images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 11–19, 2017.

Tobler, W. R. A computer movie simulating urban growth in the detroit region. *Economic geography*, 46:234–240, 1970.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. Attention is all you need. *CoRR*, abs/1706.03762, 2017. URL http://arxiv.org/abs/1706.03762.