

From Unstructured Administrative Data to a Harmonised European Reference Dataset for Machine Learning in Remote Sensing

Maja Angela Schneider

Vollständiger Abdruck der von der TUM School of Engineering and Design der Technischen Universität München zur Erlangung des akademischen Grades einer
Doktorin der Naturwissenschaften (Dr. rer. nat.)
genehmigten Dissertation.

Vorsitz: Prof. Dr. Katharina Anders

Prüfende der Dissertation:

1. Prof. Dr. Marco Körner
2. Prof. Dr. Almut Veraart

Die Dissertation wurde am 05.08.2024 bei der Technischen Universität München eingereicht
und durch die TUM School of Engineering and Design am 04.02.2025 angenommen.

Abstract

This dissertation examines the methods and findings behind the most extensive dataset used as reference data for Machine Learning (ML)-based Land Cover Classification (LCC) with remote sensing imagery.

The discovery of the secondary use of administrative data as training data for ML models is responsible for a significant leap in research, especially in the field of large-scale LCC from optical satellite imagery. Individual crop datasets collected from the member states of the European Union within the framework of the Common Agricultural Policy (CAP) can be of great scientific impact if made publicly available and harmonised across the different languages and regional crop taxonomies.

EUROCROPS (Schneider, Schelte et al. 2023) has previously been presented as a solution and, in this work, it will be put into context and its key findings and larger impact analysed. This includes not only the personal publications associated with the dataset, which show the motivation (Schneider and Körner 2022), applicability of the data as reference data for ML (Schneider and Körner 2021a) and impact that exceeded the LCC use-case (Schneider, Gackstetter et al. 2025; Schneider, Marchington et al. 2022), but also new methods and discoveries based on the work.

The importance of the research is showcased by the broad acceptance and use of the dataset, as well as the insights gained from working at a pan-European level. Hence, the work includes a section on recommendations, which can be seen as an additional result and can guide future transnational initiatives.

Kurzfassung

In dieser Dissertation werden die Methoden und Ergebnisse des größten Referenzdatensatzes untersucht, der zur Klassifizierung der Landbedeckung mit Fernerkundungsbildern mittels maschinellen Lernens verwendet werden kann.

Die Entdeckung der Sekundärnutzung von Verwaltungsdaten als Trainingsdaten für Machine Learning (ML)-Modelle hat zu einem großen Fortschritt in der Forschung geführt, insbesondere auf dem Gebiet der großflächigen Landbedeckungsklassifizierung aus optischen Satellitenbildern. Einzelne Agrardatensätze, die von den Mitgliedstaaten der Europäischen Union im Rahmen der Common Agricultural Policy (CAP) erhoben werden, können von großem wissenschaftlichem Nutzen sein, wenn sie öffentlich zugänglich gemacht und über die verschiedenen Sprachen und regionalen Agrartaxonomien hinweg harmonisiert werden.

Der EUROCROPS Datensatz (Schneider, Schelte et al. 2023) wurde bereits als Lösung vorgestellt. In dieser Arbeit wird er in einen größeren Kontext gestellt und die wichtigsten Ergebnisse und größeren Auswirkungen analysiert. Dazu gehören nicht nur die mit dem Datensatz verbundenen persönlichen Veröffentlichungen, die die Motivation (Schneider and Körner 2022), die Anwendbarkeit der Daten als Referenzdaten für ML (Schneider and Körner 2021a) und die über die Anwendung für Landbedeckungsklassifizierung hinausgehenden Auswirkungen (Schneider, Gackstetter et al. 2025; Schneider, Marchington et al. 2022) zeigen, sondern auch neue Methoden und Entdeckungen, die auf der Arbeit basieren.

Die Bedeutung der Forschung zeigt sich in der breiten Akzeptanz und Nutzung des Datensatzes sowie in den Erkenntnissen, die durch die Arbeit auf gesamteuropäischer Ebene gewonnen wurden. Daher enthält die Arbeit einen Abschnitt mit Empfehlungen, welche als zusätzliches Ergebnis angesehen werden und als Leitfaden für künftige transnationale Initiativen dienen können.

Contents

| | | |
|----------|---|-----------|
| 1 | Introduction to machine learning based land cover classification with administrative data and satellite imagery | 1 |
| 1.1 | Administrative data | 2 |
| 1.2 | Satellite imagery for land cover classification | 3 |
| 1.3 | Land cover/land use classification using machine learning | 4 |
| 2 | Agricultural subsidy control data - collection and harmonisation across the European Union | 9 |
| 2.1 | Data acquired within the scope of the European Common Agricultural Policy (CAP) | 9 |
| 2.2 | Data collection across the European Union | 11 |
| 2.3 | Pan-European data harmonisation | 11 |
| 2.4 | EUROCROPS | 12 |
| 2.4.1 | Origin and development | 12 |
| 2.4.2 | Scientific problem | 15 |
| 2.4.3 | Data collection, sources and preprocessing | 17 |
| 2.4.4 | Key findings | 18 |
| 3 | Using EUROCROPS as reference data for cropland classification from space | 19 |
| 4 | Impact of contributions and recommendations | 23 |
| 4.1 | Impact of contributions and researchers opinions | 23 |
| 4.2 | Recommendations | 24 |
| 5 | Conclusion and outlook | 29 |
| 6 | Summary of publications | 31 |
| 6.1 | EUROCROPS: The Largest Harmonized Open Crop Dataset Across the European Union | 32 |
| 6.2 | Harnessing Administrative Data Inventories to Create a Reliable Transnational Reference Database for Crop Type Monitoring | 33 |
| 6.3 | [Re] Satellite Image Time Series Classification with Pixel-Set Encoders and Temporal Self-Attention | 34 |
| 6.4 | Challenges and Opportunities of Large Transnational Datasets: A Case Study on European Administrative Crop Data | 35 |
| 6.5 | Advancing Transnational Assessments of Biodiversity Drivers in European Agriculture with an Updated Hierarchical Crop and Agriculture Taxonomy (HCAT) | 36 |
| | List of abbreviations | 37 |
| | Personal Publications | 39 |
| | Bibliography | 41 |
| | Appendix | 47 |

1 Introduction to machine learning based land cover classification with administrative data and satellite imagery

All those who have had the privilege to be on a flight with a sought-after window seat while crossing over the landscapes of central Europe might have had the chance to look down on a cloudless day, to enjoy a view akin to that in Figure 1.1. Small, angled patches cover everything the eye can see, patches that provide for everyone. No matter the season, there exists a patchwork of green, orange and brown shades, making some wonder what could be growing down there at this very moment.

Despite being much higher than a commercial plane, satellites aim to capture the same view a traveller might have with the advantage of revisiting the same area regularly, seeing crop growth development and, therefore, colour changes over time. Additionally, their sensors capture more wavelengths than the human eye, making agricultural areas an unimaginable multispectral painting. With the proper knowledge and tools, these captured patterns can reveal the answer a curious person might have asked about the crops being cultivated. However, considering the size of agricultural areas, not just on a European horizon but worldwide, the amount of data that needs to be processed far exceeds a human expert's capabilities.

There exists value in putting effort into the extraction of information from this data: Predictions for crop growth, yield and stress can be derived, crop distributions extracted and decisions about fertilisers, harvesting and crop rotations made. Fortunately, automated, computer-based analyses have taken over from the human hand, making it possible to extract the relevant information by training Machine Learning (ML) models to learn from reference data corresponding between a particular satellite spectrum and a crop type. Yet, the bottleneck in this development is the amount and quality of the reference data. This work will focus on the background and scientific questions behind developing a large pan-European crop dataset, which can help develop methods at the relevant scale and understand the principles underpinning the domain.

The key to the development of the dataset is the discovery of administrative data as a basis for scientific models; a brief overview will be given in the next section, followed by a more detailed introduction to the satellites commonly used to capture agricultural developments. Relevant work for the classification of Earth's surface with ML will motivate working with the relatively complicated agricultural administrative data across Europe, as there have been promising developments on smaller scales. The second chapter will proceed to fully explain the background of the administrative crop data itself and its path to becoming the harmonised European dataset now referred to as EUROCRIPS. Two chapters about its impact follow: one describing cropland classification and another about the broader impact of the work and recommendations. By the final chapter, the fundamentals and justifications underpinning this work have been established and, by taking recommendations into account, future pan-European research can be conducted.

This work advocates for the importance of open data and initiatives pushing for it. The challenges faced will be highlighted and examined, but the overwhelming acceptance and underestimated need for these types of datasets evidently show the asset of such an endeavour.



Figure 1.1: Agricultural fields on a clear day seen just before touching down at Munich Airport, south Germany. Source: Personal Photograph.

1.1 Administrative data

The key distinction between statistical and administrative data lies in the principal goal of the data's collection: Statistical data is acquired explicitly for statistical purposes, while administrative data is simply one of the outputs of operational systems with any statistical analysis performed as a secondary objective (Nordbotten 2010). This is clear in behavioural sciences, where administrative data has been a basis for research for years. The low costs, high volume, and unbiased nature, alongside non-invasive methods of obtaining the data, make it a prime candidate for research if handled with sufficient care and expertise (Yampolskaya 2017), and notably so as reference data for machine learning systems. This is furthered in social sciences, where administrative data has formed a cornerstone of the Big Data revolution: Its indirect nature of collection allows administrative data to fall into a subcategory of Big Data (Connelly et al. 2016). Researchers view such data with great potential despite the inherent challenges that come alongside it.

A by-product of administrative data has been the ability to perform statistical analysis on datasets initially gathered for alternate reasons, leading to the rapid development of models designed to be applied to datasets of such substantial size. However, the difficulties attributed to data quality then arise, as the original system for the collection was not designed with the intricacies in mind for applying statistical methods, but rather to run an organisation (Hand et al. 2018). Similarly reflected are the ethical and regulatory considerations of such data, which is often not made publicly available and must adhere to local data collection and distribution laws to consider General Data Protection Regulation (GDPR) concerns (Goerge 2018). Access to the data then typically requires partnering with the data providers to ensure a reasonable exchange of information. Providing this access can lead to consequential benefits, such as tackling the issue of understaffed governmental institutions that may not necessarily have the resources or funding to analyse their data. Third-party researchers can, instead, shed light on processes and drive innovation, revealing the worth of this data to the public

body. Collaboration between researchers and governments is essential for the efficient and effective use of administrative data.

One kind of administrative data, which is the basis for this thesis, is the European agricultural subsidy data. This type of data is collected within the Common Agricultural Policy (CAP), where farmers must submit self-declarations of their cultivated crop on a parcel level as a control instance for subsidy recognition, as further explained in Chapter 2.1. Some countries within the European Union (EU) have made the self-declarations publicly available. While not intentionally gathered to foster research on a pan-European scale, researchers have benefited from increasingly rich reference data for annotating satellite imagery.

1.2 Satellite imagery for land cover classification

Satellite Remote Sensing (RS) imagery, primarily that obtained by optical sensors, builds the foundation for large-scale Land Cover Classification (LCC). In recent years, increasing numbers of RS satellites have been launched into orbit, offering a range of resolutions and use-case-dependent sensor properties. LCC, however, needs a trade-off between captured area, revisit time, spectral bands, cost and many other properties, with each application potentially varying in requirements. The two most prominent programs, introduced below, the Copernicus Sentinel-2 and Landsat missions, offer both a satellite constellation with low revisit time, worldwide coverage and free data access, making them suitable candidates for large-scale analyses.

Copernicus Sentinel-2 One of the main objectives of the Sentinel-2 satellite constellation¹ is to provide open data to support the monitoring of the Earth, notably land management, agriculture and forestry. The multispectral instruments of the twin satellites provide data with a spatial resolution of 10m to 60m in the spectral range of 400nm to 2200nm in 13 bands, as shown in Figure 1.2. The wide acceptance of the data in the research community (~130.000 articles on Google Scholar, ~12.000 articles on Scopus) is backed by the short revisit time of five days and the free access to the data. Despite all the advantages, data from Sentinel-2 still suffers from cloud coverage, like any other optical sensor. This is commonly addressed by fusing the image data with the Synthetic-Aperture Radar (SAR) data from Sentinel-1 (Ibrahim et al. 2023; Schmitt, Hughes et al. 2019), but will not be discussed in depth in this work. Most prominent applications for Sentinel-2 include Land Cover/Land Use (LCLU) classification with supervised, unsupervised, pixel- and object-based methods, monitoring of forests, carbon, urban and agricultural areas as well as natural hazards (Phiri, Simwanda et al. 2020).

Landsat Whereas the Sentinel programme is a relatively new initiative, the first Landsat² satellite was launched in 1972 and offers thermal data since Landsat 4, in addition to optical. There are several satellites in use, with the most recent one - Landsat 9 - active since 2021 and providing 11 spectral bands in the range of 400nm to 2300nm, similar to its predecessor Landsat 8, illustrated in Figure 1.2. The spatial resolution is 30m for all bands, apart from the panchromatic with a 15m resolution, and the revisit time is eight days. Data can also be accessed for free, and the extensive and continuous archive provides researchers with records for long-term analysis and monitoring of the Earth's surface. Due to the mission's extended runtime, LCC methods that employed Landsat data range from early visual classification to the now prevailing ML methods (Phiri and Morgenroth 2017). With these, Land Cover (LC) classes such as urban and agricultural areas and open and dense forests could be classified.

¹ <https://sentinels.copernicus.eu/web/sentinel/missions/sentinel-2>

² <https://landsat.gsfc.nasa.gov/>

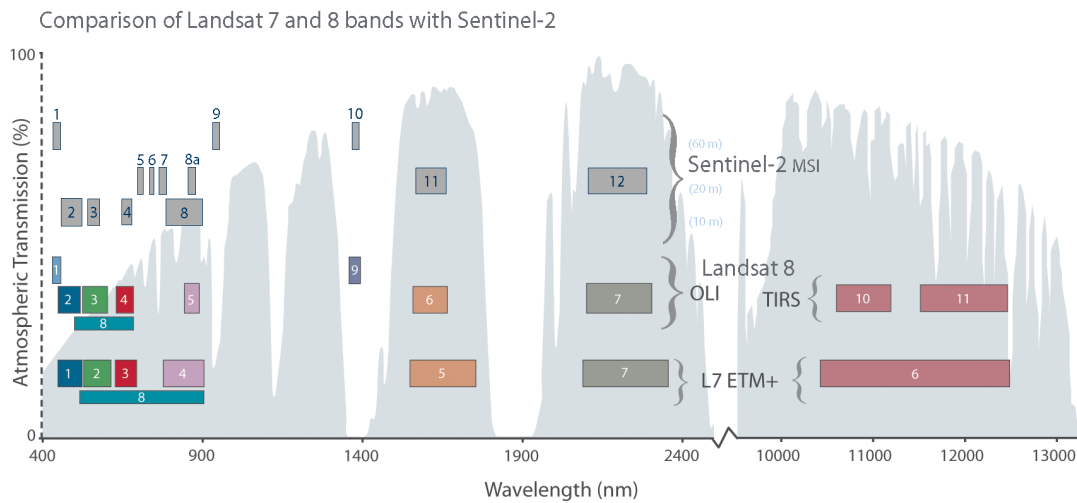


Figure 1.2: This schema visualises the different bands Sentinel-2 (S2) and Landsat (L) utilise, as well as the atmospheric transmission. As both constellations attend to similar use cases, the optical bands are generally consistent, namely: the aerosol band 1; the blue, red and green bands at 2, 3, 4; near-infrared (NI) at 8 (S2) and 5 (L); short wavelength infrared 1 (SWIR1) the bands 11 (S2) and 6 (L); SWIR2 at 12 (S2) and 7 (L); as well as the cirrus band 10 (S2) and 9 (L). While Landsat additionally offers the high-resolution panchromatic band 8 and thermal bands, Sentinel-2 has more channels in the red edge/NI and generally higher spatial resolution on most bands. Source: <https://landsat.gsfc.nasa.gov/wp-content/uploads/2015/06/Landsat.v.Sentinel-2.png> and <https://www.usgs.gov/media/images/comparison-landsat-7-and-8-bands-sentinel-2>

Other missions Sentinel-2 and Landsat cover many applications and offer a usable trade-off between spatial and spectral resolution, as well as cost and revisit time. However, there are other missions that are contributing now or in the near future to the field of RS LCC. In 2022, the German Aerospace Centre (DLR) successfully launched the hyperspectral EnMAP³ satellite that offers 246 spectral bands and a spatial resolution of 30m, but at the cost of a lower a revisit time of 27 days. Beyond that, commercial satellite missions like WorldView⁴ and the constellation from companies like Planet⁵ provide higher resolutions, both spectral and spatial, but are not designed for large scale LC mapping and usually require proposals or funds in order to to get access.

Altogether, it becomes evident that the mass of data collected daily by fleets of satellites depends upon stringent mechanisms and automation to process and analyse the material. Researchers have therefore taken the next step to explore and justify the feasibility of ML methods for LCC. However, the limited amount of reference data for most use cases restricts the analysis to a spatial extent or in a way that the outcomes are too coarse to provide impactful scientific results.

1.3 Land cover/land use classification using machine learning

One of satellite imagery's most prominent use cases is mapping Earth's surface, most notably, the classification of landmass into different categories. This can range from a very coarse differentiation between agricultural, urban or forestry areas, which is usually referred to as Land Cover Classification or mapping, to the finer so-called Land Use (LU) classification or mapping, which includes classes

³ <https://www.enmap.org/>

⁴ <https://www.maxar.com/maxar-intelligence/constellation>

⁵ <https://www.planet.com/products/monitoring>



Figure 1.3: An Land Use (LU) map, in this case for crop classes, visualises the different geo-referenced field parcels with the corresponding categories by distinct colours. Here, the map overlays a Sentinel-2 RGB image, making it easy to spot urban and forest areas that are not part of a crop map. The image was generated using administrative data as reference data, but parcel-based classification algorithms can lead to similar-looking results. Image Source: Schneider and Körner (2022)

like specific building or crop types, as shown in Figure 1.3. Here, the advantage of having a relatively short revisit time of the previously introduced satellites comes into play: by analysing a Satellite Image Time Series (SITS), the spectral development of a specific application can be examined and higher accuracies achieved. Crops, for example, reflect different wavelengths over time as they grow and, by using several consecutive images, they can be classified far more easily than with just a single frame. The resulting maps can help retrieve information about the studied areas and thus support decision-making processes and scientific progression.

Early methods With the ongoing increasing quantity of Earth Observation (EO) satellites, the amount of imagery data that needs to be examined and mapped grows exponentially. Traditionally, RS experts examined the data acquired from the satellite and utilised a Geographic Information System (GIS) to apply hand-crafted feature extraction to obtain information (Rawat et al. 2013). In parallel, statistical methods like Maximum Likelihood Estimation (MLE) and Principal Component Analysis (PCA), used to assess satellite images to derive information about land cover, were developed (Aroma and Raimond 2016). Soon, it became evident that too much data was being gathered and the processes needed to deal with this amount had to scale without the laborious employment of expert knowledge on an image basis and the use of compute-intensive algorithms. As a counteract, indices like the Normalized Difference Vegetation Index (NDVI), which combined several satellite bands, became a newly standardised tool to classify large areas simultaneously (Foerster et al. 2012; Wardlow and Egbert 2008), as well as ML methods like Random Forests (RFs) and Support Vector Machines (SVMs). RFs were used to assess the feasibility of Sentinel-2 data for crop type mapping before the mission had become operational (Inglada et al. 2015), as well as the accuracy of multi-year crop type mapping (Vuolo et al. 2018) during the first years of service. Similarly, Kang et al. (2018) employed an

SVM in the early years of Sentinel-2 to showcase the potential use and application of ML models for crop mapping. All this showed that creating LCLU and crop maps from satellite images with ML was achievable and could be used to assess the Earth's surface regarding agricultural and natural monitoring, climate impacts, and many more applications. But the development of data-driven algorithms did not stop there: The prevailing trend towards Deep Learning (DL) also captured the community of EO and RS scientists. Esteves and Valente (2024) compared an SVM, an RF, and a Convolutional Neural Network (CNN), namely the U-Net, and concluded that, for a task like country-wide LCLU mapping, DL architectures exceeds the performance of known ML methods.

The rise of Deep Learning methods RFs and SVMs were responsible for a leap in accuracy and have made their way into the standard toolbox of an RS image analyst, but the rise of DL kicked off an entirely new dynamic in the field. Suddenly, large amounts of data could be more accurately processed. The only bottleneck in the pipeline was the lack of reference data, requiring researchers to step back and employ weakly supervised methods (Schmitt, Prexl et al. 2020). Other approaches required the use of dedicated test sites, for example in Ukraine, where Kussul et al. (2017) assessed the performance of a CNN in comparison to Multilayer Perceptron (MLP) and RF baselines, outperforming both. The trend towards modern DL approaches was then shown in a wide range of experiments, where other types of reference data, such as LUCAS (d'Andrimont et al. 2022), were also used to assess the performance of a network consisting of an Recurrent Neural Network (RNN) and a CNN (Mazzia et al. 2020), again beating an SVM and RF. However, the lack of dedicated reference data that could elevate the proven performances of DL methods still commanded the developments.

Higher accuracies with administrative data The discovery of administrative crop data, described in more detail in Chapter 2.1, as reference data for SITS was the ignition of research in the field. In the beginning, only smaller sectors allowed for the analysis of RNNs, CNNs, and SVMs covering larger areas, as demonstrated in the region of Bavaria in Germany (Rußwurm and Körner 2017). The data collected there sparked the development of new semantic segmentation models with a U-Net and C-Long Short-Term Memory (LSTM) (Rustowicz et al. 2019). This showed that, despite not primarily being collected for the purpose of LCLU mapping of SITS data, researchers finally found a means to test their theoretical approaches on a relevant amount of data. Rußwurm and Körner (2018) continued their research and dove deeper into the findings of their earlier work, where they discovered that modern DL architectures, such as ConvRNN, can efficiently handle cloudy and noisy data and are therefore able to use unprocessed Top Of Atmosphere (TOA) Sentinel-2 data.

The utilisation of large-scale French administrative crop data built the new foundation for assessing common DL methods on an even greater magnitude. Garnot et al. (2019) applied a CNN, Gated Recurrent Unit (GRU), and two hybrid approaches of an R-CNN to the data. With their detailed study on the data and the methods, they set a new baseline for new approaches to compare their work against. In a concluding remark, they further mentioned that an RF baseline was easily outperformed, highlighting the clear transition in research from traditional ML models to DL. Pelletier et al. (2019) used similar data and ran a comparative study on TempCNNs and RNNs, where the former outperformed the latter, as well as an RF. While they used analogue data over Sentinel-2, they discovered that using vegetation indices like NDVI did not improve the performance of a Neural Network (NN). Researchers working with Greek agricultural reference data came to the same conclusion that DL outperforms traditional ML procedures for LC mapping with Sentinel-2 (Papadopoulou et al. 2023). By utilising Swiss data, Turkoglu et al. (2021) were able to further push the boundaries of commonly used methods by introducing a hierarchical convRNN that successfully classified the pixels in a given Sentinel-2 patch. While focusing on different aspects of the application or methodology, all these studies conclude that employing larger-scale training datasets can widen the performance gap between the various ways to analyse satellite imagery, and new and better methods for reliably extracting information could be

designed.

When finally Transformers (Vaswani et al. 2017) came into the spotlight, Rußwurm and Körner (2020) conducted a first study for employing the self-attention mechanism for SITS data, determining that the inherent structure of the architecture, similarly to LSTM-RNN, allows better management of the noisy and unprocessed Sentinel-2 time series data. The most recent leap in the field of ML-based LCC was by Garnot et al. (2020), introducing a combination of a Pixel-Set Encoder (PSE) and a Temporal Attention Encoder (TAE), based on the Transformer architecture, both of which pushed the boundaries of accuracies and made it a standard tool, even outside academia (Barrett and Toro 2024). Their research was made possible due to their access to vast quantities of crop reference data within France.

2 Agricultural subsidy control data - collection and harmonisation across the European Union

The previous chapter introduced the possibilities of data-driven methods to automatically process satellite imagery for LCLU. It is well established that these algorithms are incredibly hungry for reference data; hand-made annotations have always lacked that capacity. This has led to a shift in focus towards using administrative data that has already been collected, holding relatively high quality and relevant quantities. Now, these two observations will be bound together: Researchers need reference data, and an immense pool of hardly used administrative data exists. Schneider and Körner (2022) motivated that bridge extensively and illuminated the fit of administrative data inventories for research and innovation. While the data is usually not standardised and due to different data formats not interoperable, the authors see great benefits in the high granularity, wide coverage, and regular collection. Large transnational datasets hold the potential for cross-validation and quality improvement of individual datasets, and the great extent of data enables spatial and temporal analysis.

The core of this thesis and presented solution to the aforementioned requirements is the EURO-CROPS dataset, which will be introduced and explored in the following sections. Rooted in the CAP of the EU, the data and its relevant components are first defined, yielding a solid foundation for the subsequent data collection and harmonisation sections. Lastly, the origin, scientific problem, methods and key findings associated with the EURO-CROPS dataset are presented. There, complementary information to the publication by Schneider, Schelte et al. (2023) is given, completing the story of the data.

2.1 Data acquired within the scope of the European Common Agricultural Policy (CAP)

Chapter 1.3 previously introduced agricultural administrative data as reference data for ML. However, its origin and background have not yet been expanded on. This section will introduce and explain the technical terms, providing a basis for the upcoming central part of this work.

Within the EU budget, the CAP receives over 300 billion Euro¹, making it one of the biggest beneficiary of funds. Member States (MSs) are responsible for ensuring the accurate distribution of subsidies and, therefore, must collect and process application data. There are three terms connected to that process, IACS, LPIS and GSA, that all play a role in the development of pan-European agricultural datasets.

Integrated Administration and Control System (IACS) MSs or, in some cases, individual regions, ensure the correct implementation of CAP via the so-called Integrated Administration and Control System (IACS) (Martirano and Toth 2023). These hold the information for putting CAP nationally or regionally into place. There are around 40 IACS systems in the EU (Van der Velde 2021), with each MS or region being allowed to decide how to implement it and perform quality assurance. Within the system, two sets of data are needed to ensure the correct distribution of subsidies: LPIS and GSA.

Land Parcel Identification System (LPIS) The first subset of data within IACS is the LPIS, geo-spatial records that define the field boundaries and land cover properties, indicating whether a parcel is eligible for subsidies. In Figure 2.1, the parts of the data that are collected and updated in regular intervals as part of LPIS are depicted in red. It is recommended to be referred to as land cover (Toth

¹ https://agriculture.ec.europa.eu/common-agricultural-policy/financing-cap/cap-funds_en

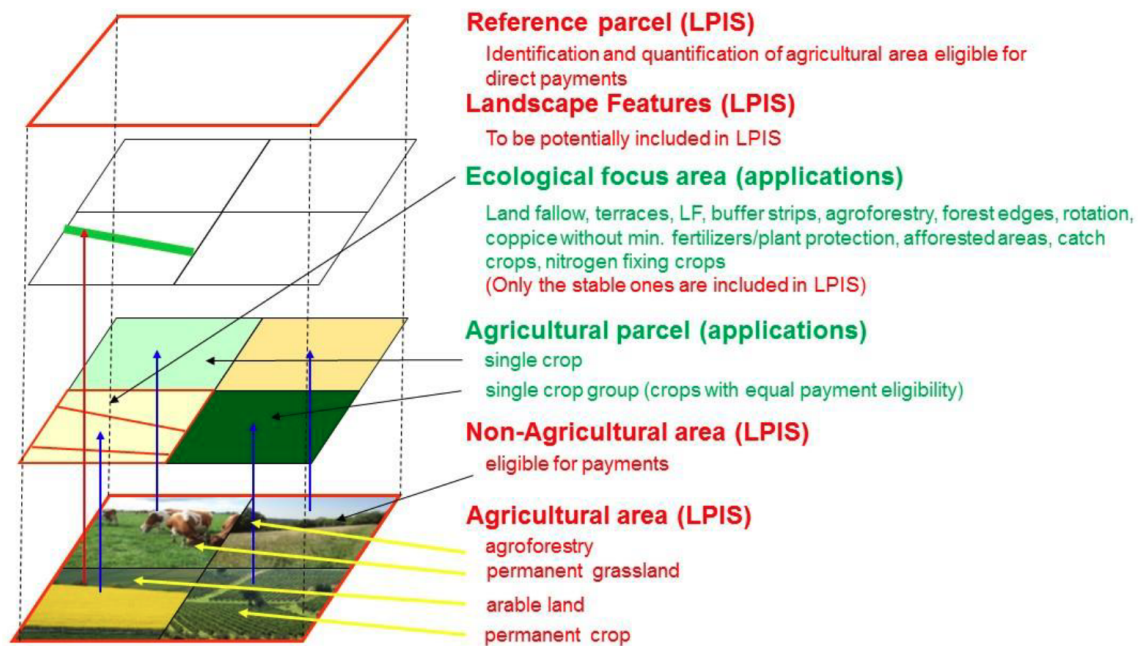


Figure 2.1: The two components of an IACS, LPIS and GSA, are visualised together. From top to bottom, this first shows the general geo-referenced parcel eligible for subsidies. This can either be an ecological focus area, an agricultural parcel (both recorded within GSA) or a non-agricultural area (recorded in LPIS), which is still eligible. While the LPIS parts of the data, indicated in red, show general properties such as arable land and permanent grassland or crop, the green GSA, or here called "applications", parts provide finer distinctions, such as the crop type. Source: Martirano and Toth (2023)

and Milenov 2020) and describes properties as agricultural and non-agricultural areas, landscape features and reference parcels.

Geospatial Application (GSA) On the other hand, and the integral part of the previously mentioned agricultural datasets, MSs need to collect GSA data, formerly known as Geospatial Aid Application (GSAA) (Martirano and Toth 2023). This collection holds the yearly agricultural application data of the farmers in the EU. In particular, the crop is cultivated for each parcel, defined by geo-referenced LPIS field delineations. The crop data is referred to as Land Use (Toth and Milenov 2020) and is indicated in green in Figure 2.1.

Impact on research Researchers have understood the potential of both the LPIS and the GSA data for a long time for applications like data quality control and refining LCLU information and their respective zones and typologies (Baiamonte et al. 2023). In recent years, many datasets, applications and methodologies using such data have been developed. This led the European Commission (EC) to decide that going forward, the agricultural administrative data will be referred to as a high-value dataset and, therefore, mandatory to make publicly available (European Commission 2022).

The impact on the development of new ML models has already been highlighted in Chapter 1.3, with this type of data being the answer to the call for training resources. Prominent datasets, such as BreizhCrops (Rußwurm et al. 2020) and ZueriCrop (Turkoglu et al. 2021) are just two examples of the initiatives for making use of the data in the research sector. However, most uses are limited to single countries due to a missing pan-European data collection and harmonisation initiative. EURO CROPS will be presented in Chapter 2.4 as a solution, but first, the next two sections look into research and how data in Europe can be collected and harmonised.

2.2 Data collection across the European Union

As mentioned, LPIS and GSA databases are set up and maintained regionally or nationally. This setting requires planning and executing a structured method to locate all data sources, contact the data providers, and obtain different collections. One way to plan out this activity is to look for similar workflows in other domains and adapt them to the given needs. In this context, one example is the European exchange and collection of health data: Here, research networks provide the means to gather information from administrative sources, health surveys and studies. Unim et al. (2022) describe that health data within the EU is distributed in more than 57 thematically and partially overlapping research networks, making it accessible but often fragmented across different locations. This shows that there are procedures in place that enable researchers to benefit from a significant accumulation of national datasets.

In the case of LPIS/GSA data, recommended by the EC, the appropriate platform for the MSs to publish the data would have been the Infrastructure for Spatial Information in Europe (INSPIRE)² and especially the INSPIRE Geoportal (Toth and Milenov 2020). Here, MSs can publish and share their data and services and, together with the platform providers, contribute to data that follows the FAIR principles³ (*Findability, Accessibility, Interoperability, and Reusability*). However, during the first steps towards EUROCROPS, only two countries (See Table 4 in Schneider, Schelte et al. (2023)) published their datasets on the portal, highlighting the difficulties for researchers if the principles are not followed.

Additionally, national data protection regulations play a significant role in collecting information across countries. Based on the GDPR, MSs can build their regulatory framework on how to implement legislation on their data. This made it increasingly complex to convince MSs to publish their datasets as some countries draw the line at different points regarding administrative information. So even though LPIS and GSA are part of the official database of high-value datasets, some countries' regulations might have conflicted with that, as it depends on the government what is being seen as personal data and therefore be protected by the GDPR (van Loenen et al. 2016).

2.3 Pan-European data harmonisation

Once data is collected, stored and checked, interoperability needs to be ensured by deciding on the common ground for all sub-datasets. In the pan-European case, this includes dealing not only with technical differences but also with language, culture, and law. To use several individual data collections as one common basis for research and policy, data has to be transformed into a format with attributes that allow direct comparisons between granular entries of different sources. Generally, this can either be achieved by defining a mapping from one source to another or by developing a new schema in which all sub-collections are aligned.

Data harmonisation as a research subject The BioSHaRE Project (Doiron et al. 2013) is one of the most prominent examples of building a transnational harmonisation framework. For the study, data from six countries build the basis for algorithms that automatically transform the individual collections into a target format and store them in distributed servers across Europe. It shows that cross-border research is possible, efficient, and secure while also stressing the importance of conducting regular meetings and workshops with all participating parties to identify the correct target variables.

Data harmonisation is an investigated topic in research, further demonstrated in health data, where there has been a push towards conducting the process in a standardised way (Gurugubelli et al. 2022). This includes developing data harmonisation plans and protocols, commonly defined variables

² https://knowledge-base.inspire.ec.europa.eu/index_en

³ <https://www.go-fair.org/fair-principles/>

and dictionaries, and a dedicated way of storing the data on distributed servers. Additionally, communication between parties to enable understanding and transparency regarding common goals is one of the most crucial aspects of a common basis of data harmonisation. As a result, access to and usability of transnational datasets enables European health research to tackle far more complicated problems than previously achievable.

Taking research into practice These lessons can also be applied to the present case of LPIS/GSA data. The previous chapters have already established the availability of, or at least the push towards, making this type of data open. But the rich diversity of the MSs of the EU makes it difficult to establish one system. Therefore, data in the current environment must be harmonised post-collection, and finding common ground and mappings between individual datasets becomes incredibly challenging due to the practically unknown number of stakeholders.

One method in which to approach this problem is through the use of a minimal harmonisation profile (Stoter et al. 2022) where a partial harmonisation of the data is conducted and results in an interoperable dataset that includes sub-datasets from different sources or types. By finding a minimal set of variables that build the basis for common ground, trust between parties can be established and strengthened.

In the case of EUROCRIPS, the first common ground was the EAGLE matrix (Arnold et al. 2013) provided by the European Environment Agency (EEA). During the pilot phase of the project (Schneider, Broszeit et al. 2021), data was harmonised across a very small number of classes. Despite the limited expressiveness, the trust of researchers was gained and the need for such a dataset was substantiated.

Altogether, there are three non-exclusive main challenges that data harmonisation across country borders needs to tackle:

1. Diverse standards and heterogeneity of dataset contents and formats lead to bottlenecks in the process (Gu et al. 2021). Differences between academia and industry add to a drift between the goals of standardisation.
2. Legislative regulations of MSs, which result in different parts of data being published and withheld.
3. Trust of the population and MSs that data sharing and harmonisation benefit society.

2.4 EUROCRIPS

The first full harmonised dataset version⁴ of EUROCRIPS (Schneider, Schelte et al. 2023) was published in April 2022 on Zenodo (Schneider, Chan et al. 2023). It followed a year-long manual collection, processing and verification process that had built on top of the pilot project TINYEUROCRIPS (Schneider, Broszeit et al. 2021; Schneider and Körner 2021b).

The previous sections already introduced the type of data that holds the basis of the research project, followed by general introductions to working with pan-European data by first collecting and then harmonising it. Now, as this foundation is laid out, the focus will be on the background and challenges of the curation and development process of the dataset.

2.4.1 Origin and development

Pilot: TINYEUROCRIPS As a consequence of the developments mentioned in Chapter 1.3 and the increasing availability of data and data processing structures across the EU, EUROCRIPS was

⁴ Version 1: <https://zenodo.org/records/6866847>

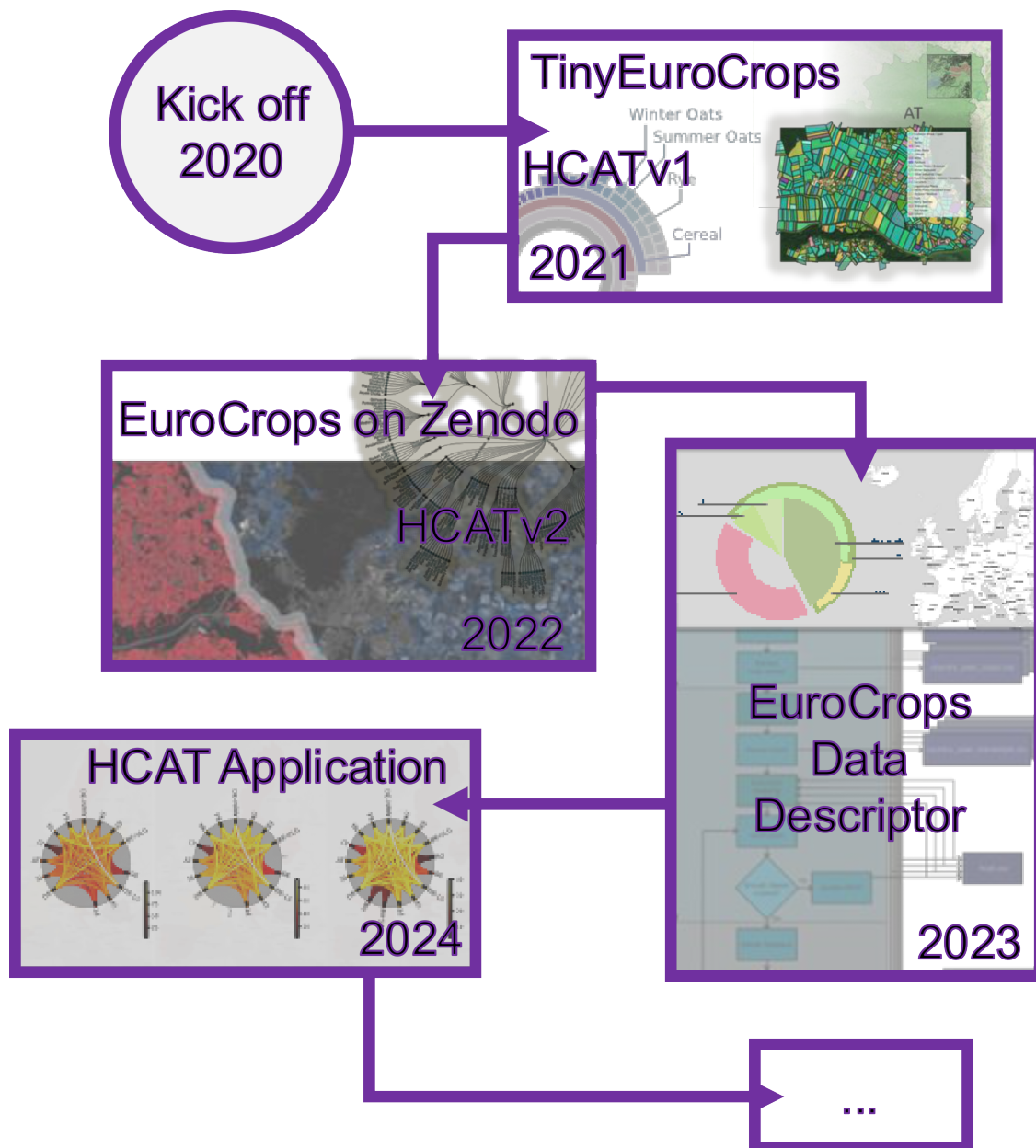


Figure 2.2: This scheme summarises the development of EUROCROPS over time. The first pilot kicked off in 2020 and resulted in 2021 in the demo dataset TINYEUROCROPS with the corresponding first version of Hierarchical Crop and Agriculture Taxonomy (HCAT). It only included data from Austria, Slovenia and Denmark but showed the potential of pan-European data collection and harmonisation. In 2022, the first release of an updated HCAT together with harmonised vector data from a large number of MSs of the EU was uploaded to Zenodo, followed by the official data descriptor being published in 2023 in Scientific Data. All harmonisation processes, data sources, and background information were brought together and explained. In 2024, a large pan-European study on biodiversity drivers was conducted with the help of HCAT and showed the impact of the work beyond the originally intended disciplines. Image Sources: 2021: Schneider, Broszeit et al. (2021), Schneider and Körner (2022), 2022: Schneider, Gackstetter et al. (2025) and <https://github.com/maja601/EuroCrops> and <https://www.eurocrops.tum.de/taxonomy.html>, 2023: Schneider, Schelte et al. (2023), 2024: Schneider, Gackstetter et al. (2025)

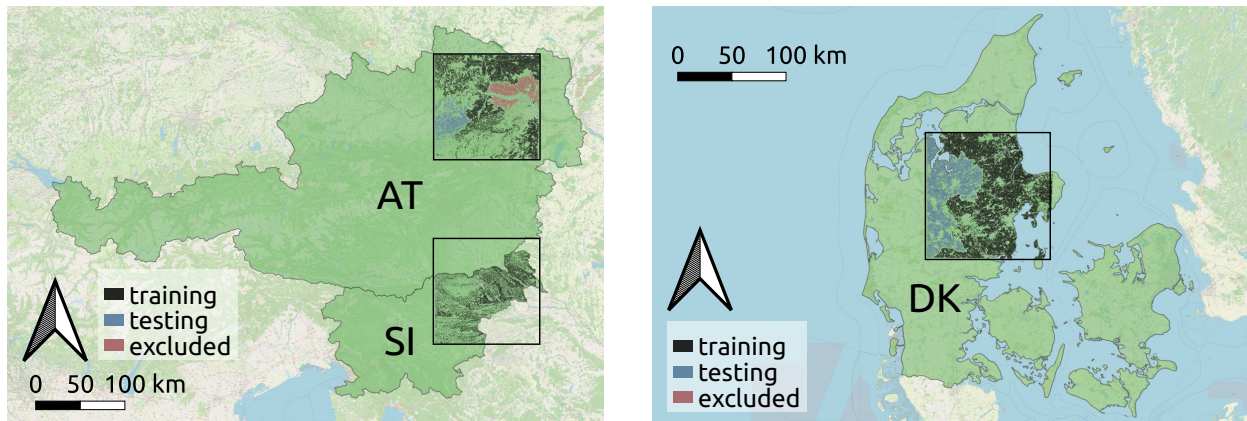


Figure 2.3: TINYEUROCROPS consists of three patches with harmonised reference data and the corresponding Sentinel-2 SITS reflectance data for developing ML models. The chosen areas are Austria, Slovenia, and Denmark, which enable cross-country validation with similar and different climate conditions. The data is already spatially divided into train and test areas, as well as some excluded areas for potential benchmarking. The dataset is available on MediaTUM (Schneider and Körner 2021b). Image Source: Schneider, Broszeit et al. (2021)

launched as a proposal to address the issues related to the challenges that researchers in the domain faced. Initially, the idea was to investigate the possibilities of utilising larger scale CAP data for Artificial Intelligence (AI) applications. In collaboration with the company GAF AG⁵, the first spreadsheets for data sources and crop classes were set up. While the hunting for public data took several months, the decision to base the taxonomy for the crops on an early version of the EAGLE matrix (Arnold et al. 2013) was made relatively early due to its comparatively simple and clear structure for crop types. It was decided that only a few classes and seasonality for the main cereal types were added to the original design. Although the new taxonomy proposal only included the acquisition and harmonisation of data from three countries, the curiosity of how much data could potentially be available triumphed: it was possible to get in touch with ministries or identify publicly available data sources from thirteen countries. However, for the first iteration of the initiative, only data from Denmark, Austria and Slovenia was translated, as depicted in Figure 2.3, and harmonised with the simpler but newly introduced first version of Hierarchical Crop and Agriculture Taxonomy (HCAT). As the AI application narrative was still driving the development, the vector data in three areas was enriched with Sentinel-2 SITS data to enable researchers to apply their crop type classification models straightforwardly.

The publication of the first EUROCROPS dataset, published on MediaTUM⁶ and later also referred to as TINYEUROCROPS (Schneider, Broszeit et al. 2021; Schneider and Körner 2021b), was the critical milestone that lit the spark of realisation for how beneficial a data collection exercise could be. Surprisingly, researchers and industry took much higher advantage of the mappings between the country-dependent crop taxonomies and HCAT instead of the prepared crop classification dataset with SITS. It enabled the usage of vector data as labels for their imagery- and country-agnostic research, which is further expanded on in Chapter 3. Another insight was that while most publications focused on the most common crops or grassland, there was a particular interest in the less frequent classes.

Scaling up data and taxonomy Based on these findings, a second vector-data-focused dataset was developed without external partners. Schneider, Schelte, Schmitz and Körner worked on obtaining more data, increased the granularity of HCAT and published ‘EuroCrops: The Largest Harmonized Open Crop Dataset Across the European Union’ (Schneider, Schelte et al. 2023) as a detailed data descriptor, including all data sources and processes, with the corresponding mappings from each

⁵ <https://www.gaf.de/>

⁶ <https://mediatum.ub.tum.de/1615987>

countries' crop taxonomy available on GitHub⁷.

Alongside the data, HCAT evolved over time: what started as a slight extension to the EAGLE matrix became an elaborate taxonomy spanning several hundred classes, representing most of the crops being cultivated and declared in the EU, visualised in Figure 2.4. One clear advantage was the clear code and levels, allowing new classes to be easily fitted into the schema and collaborators able to raise GitHub Issues with recommended extensions and bugs they had discovered. Upon acceptance, merging the new addition directly into the mappings was possible, allowing both backwards compatibility and extensibility for new data.

2.4.2 Scientific problem

EUROCROPS addressed the issue of integrating and standardising diverse, large-scale datasets. They, therefore, enabled scientists to perform sophisticated and comprehensive analysis for a wide range of agricultural and RS research applications. Two main challenges had to be considered and were attempted to be solved during the development.

Analysing large-scale datasets Firstly, there were difficulties in processing and surveying extensive and varied data collections. In contrast to tabular data, geospatial data holds an entirely different magnitude of information and, consequently, that volume needs to be stored. Although vector data is the smaller of the two primary types of geospatial data, collecting and storing all publicly available field polygons of the EU requires efficient depot, retrieval and processing techniques and infrastructure. Additionally, the inherent complexity of the data and inconsistent data formats, such as a shapefile or geopackage, added to the arising problem when working on a general solution, due to the likely non-existence of a one-fits-all format.

EuroCrops-specific data challenges Secondly, data-specific issues arose during the development. The MSs of the EU still hold the sovereignty over their countries' way of implementing IACS and, therefore, most of the contents in such a data collection. The heterogeneity of crop names, codes, declaration and publishing granularities, platforms, and regulations made it incredibly difficult to standardise and prepare the data for transnational analysis. Despite all efforts in the past years, EUROCROPS does not yet cover the entirety of the EU. This is a common issue in pan-European research, but the constant push towards more open data is starting to show success. Data from some countries, such as Finland, was not accessible in earlier stages but has recently been made available. Official bodies, such as the EC, are now supporting the open data movement and declared LPIS/GSA as high-value datasets (European Commission 2022), making them mandatory to be publicly available in the near future. The biggest data-specific challenge, however, was the lack of a European-wide taxonomy that covered all MS-specific schemes and allowed for transnational crop categorisation and comparison. The EAGLE matrix was undoubtedly not the only option as a starting point, but it had the apparent benefit of being the most promising and straightforwardly extendable one. The development of HCAT subsequently required extensive collaboration with experts and representatives of the participating countries to accurately reflect the agricultural practices and crop types present in the EU. It, therefore, needed to be flexible enough to accommodate various regions, granularities and implementations of IACS. As a result and visualised in Figure 2.4, HCAT had to be detailed enough to cover the finest crop distinctions in the original data from each MS and, with a hierarchical structure, to also correctly include coarser data collections.

⁷ <https://github.com/maja601/EuroCrops>



Figure 2.4: The hierarchical structure of HCAT is visualised in a dendrogram. Each circular layer shows one stage of granularity, starting from coarse differentiation into permanent crops and arable crops in the central area to the fine seasonal distinction into winter and summer wheat in the outer band. Source: (Schneider, Gackstetter et al. 2025) and <https://www.eurocrops.tum.de/taxonomy.html>

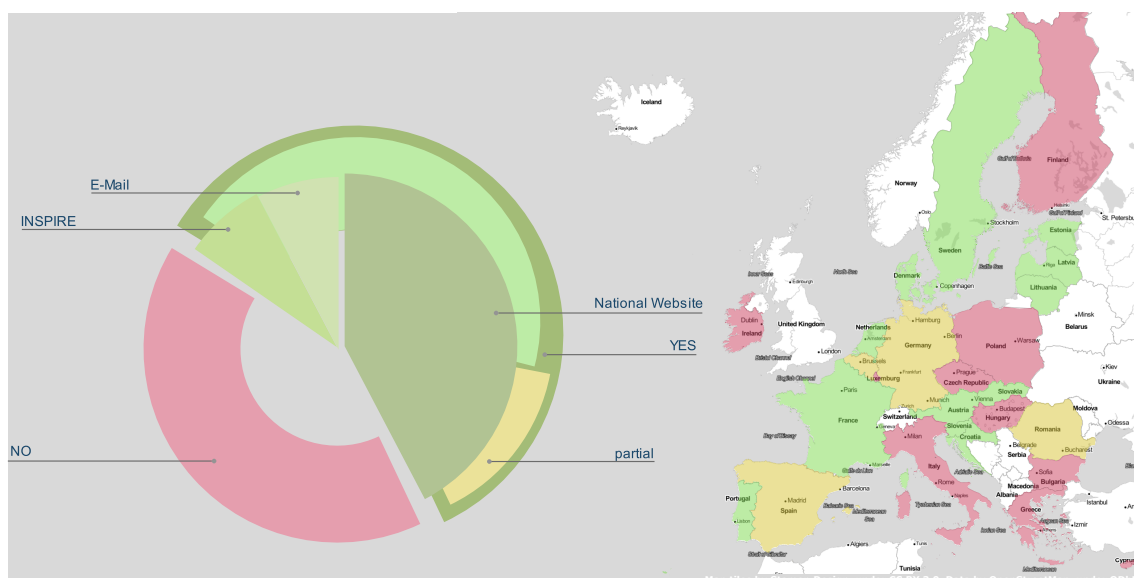


Figure 2.5: The scheme published originally in ‘EuroCrops: The Largest Harmonized Open Crop Dataset Across the European Union’ (Schneider, Schelte et al. 2023) shows the distribution of available administrative crop data across the EU. The map on the right indicates whether a country releases their data fully (green), partially (yellow) or not at all (red). The chart on the left breaks down the availability further by showing where the data from the green and orange countries was obtained.

2.4.3 Data collection, sources and preprocessing

As already introduced in Chapter 2.1, the basis for EUROCROPS is LPIS/GSA data, compiled within the CAP. The development of ML methods based on administrative data, especially in the agricultural sector, drove the idea for the initiative. The datasets for each country, or sometimes region, had to be obtained individually. All data sources are described in detail in ‘EuroCrops: The Largest Harmonized Open Crop Dataset Across the European Union’ (Schneider, Schelte et al. 2023) as well as on the GitHub Wiki⁸. Data availability and access during the first publishing phase of EUROCROPS is illustrated in Figure 2.5.

From raw data to EUROCROPS Upon gathering the raw data from the countries, several combined automated and manual data verification and preprocessing steps were put in place. Usually, the vector data was inspected in QGIS⁹ and, via checking the attribute tables, a first impression about the country-dependent crop declaration management was established. The attribute table was exported and the unique crop names were extracted. These had to be translated, whereas only about half of the automated translations were correct, making it a chore to manually check all, often with the help of Wikipedia or other crop-relevant resources. Then, a first mapping between translations and HCAT was automatically performed, followed by manually verifying and correcting the correspondences. If dominant classes of the raw dataset were not present, HCAT and all previously mapped datasets were updated. All automatic steps were performed with Python, whereas the manual ones were in Excel. Due to the size of the datasets, broken entries or missing values were ignored, leaving a slight but neglectable selection bias. By physically checking each crop name, it was aspired to keep the processing error as low as possible: subsequent experiments conducted by the Joint Research Centre (JRC) of the EC confirmed the primary success of the efforts while being able to point out

⁸ New Wiki: <https://github.com/maja601/EuroCrops-Wiki> (<https://github.com/maja601/EuroCrops/wiki> is deprecated)

⁹ www.qgis.org

discrepancies¹⁰ and even improve the mappings.

2.4.4 Key findings

After four years of work, EUROCRIPS has become a successful open, transnational data research demonstrator. The harmonised vector data is easily accessible and constantly being updated over time. Together with HCAT, EUROCRIPS is part of the tools researchers in the agricultural and RS sector know and utilise. Examples of the new methods developed and insights found will be shown in Chapter 3, with the dataset being the application for theoretical, new approaches. By enabling transnational research, HCAT also helped advancements in other study areas, such as research for biodiversity drivers across Europe (Schneider, Gackstetter et al. 2025). This field was originally not within the scope of the project but showcased a versatile application for the data, elaborated on further in Chapter 4.1, where the broader impact will be explained.

Informed decision making in a interdisciplinary consortium While it was evident how difficult the journey in creating a pan-European crop taxonomy might be, it was surprising how hard it was to tend to all the different disciplines that were engaged and interested in the project's outcomes. One example is RS scientists wanting the seasonality of a crop being an earlier group as winter wheat and winter barley corresponded to a more similar reflectance pattern in SITS than summer and winter barley. They would have put the general distinction between summer and winter cereal before branching into the detailed cereal types for each seasonality. On the other hand, agricultural scientists were interested in keeping the same crop in one group and only distinguishing between seasons on the finest scale because the differentiation between wheat and barley was more important than whether it was summer or winter wheat. All parties had different motivations and needs, and the impossibility of achieving an end goal that satisfied everyone soon became apparent. For a brief period, the project came to a near-halt as it appeared no consensus would be found. However, at the beginning of EUROCRIPS, the eventual impact of the project was not yet known, so it was decided to go for an easy compromise in the hierarchical structure. Not everyone received their preferred outcome but, notably, this issue was outweighed by the advantage of having a transnational scheme that worked. Now that HCAT is out, and despite all its possible flaws, researchers can work on it and improve it. This is reflected in the issues posted on GitHub or them developing straightforward mappings from their taxonomy to HCAT, making it possible for them to utilise the full range of data available within EUROCRIPS.

Large, versatile datasets over small task-specific ones In comparison to the initial idea of the project, the final results showed that the perceived need for such data was clearly understated. What had started as an initiative to gather some more data in order to develop better AI/ML/DL models evolved into an open data movement impacting more fields than assumed initially. The amount of parcel data currently present in EUROCRIPS has also resulted in the decision not to process the corresponding Sentinel-2 data for each field due to an exploding amount of storage that would have been required. Consequently, EUROCRIPS is now being hosted on several cloud platforms (see Chapter 4.1) in order to allow for online layering of vector and raster data for everyone individually, as usually the entire spatial extent of EUROCRIPS is not needed to answer a specific research question.

The broader impact inside and outside of the crop classification sector, as well as the recommendations based on the experiences of this work, will be presented in Chapters 3 and 4 respectively. An outlook into the future of EUROCRIPS will be given in Chapter 5.

¹⁰<https://github.com/maja601/EuroCrops/issues/9>

3 Using EUROCROPS as reference data for cropland classification from space

As previously mentioned, the work by Garnot et al. (2020) set a new standard for crop type classification from SITS using DL. After having developed the demo dataset TINYEUROCROPS (Schneider and Körner 2021b), a new question was raised: how well would the previously, only locally employed method work when applied in a transnational manner? In the reproducibility study ‘[Re] Satellite Image Time Series Classification with Pixel-Set Encoders and Temporal Self-Attention’ (Schneider and Körner 2021a), the adaptability of the method to data collected within the EUROCROPS initiative was investigated. In addition to the papers’ reference data from France, EUROCROPS data from Slovenia was prepared in the same manner, and the methodology was tested with a random regional split and a cross-dataset evaluation for both collections. This resulted in positive outcomes and verified both the method and the dataset for this use case.

By the time ‘EuroCrops: The Largest Harmonized Open Crop Dataset Across the European Union’ (Schneider, Schelte et al. 2023) was published, EUROCROPS had become a commonly utilised reference data collection for crop type classification with an increasing amount of new research ideas being developed with it and in excess of over ten thousand downloads on Zenodo (Schneider, Chan et al. 2023). The following will introduce an excerpt of these new methods, highlighting the importance of having a pan-European dataset.

Thermal Positional Encoding for SITS Data Nyborg et al. (2022) realised that one of the greatest difficulties of knowledge transfer in crop type classification is the variable growing periods of crops. They consequently proposed an alternative to traditional positional encoding, Thermal Positional Encoding (TPE), based on the thermal development of an area over time instead of the calendar days. As growing patterns are closely related, a better generalisation of crop type classification algorithms was achievable. While the authors are not directly using the reference data provided by EUROCROPS, there were synergies¹ during the development of it and their data collection.

Baseline Results for EUROCROPS Aszkowski and Kraft (2023) were the first to run a baseline comparison of different basic machine learning models on TINYEUROCROPS, indicating that the dataset had become a part of the crop type classification toolbox. They ran experiments for all 44 available classes, comparing a MLP with a SVM and RF and concluded that, while the MLP performs the best, training and testing in different countries still holds a lot of challenges for these straightforward methods.

Using spaceborne LiDAR for global maps of tall and short crops Di Tommaso et al. (2023) also employed TINYEUROCROPS as part of a global dataset used to develop and assess the applicability of LiDAR data acquired by NASA’s Global Ecosystem Dynamics Investigation (GEDI) mission for the distinction between short and tall crops, as well as the identification of the peak height of tall crops. EUROCROPS was used to verify the method in two additional parts of Europe: Austria and Slovenia. The Danish dataset is omitted due to its location outside the GEDI latitude coverage. This highlights the importance of keeping the access to EUROCROPS dynamic and allowing researchers to pick and choose the relevant sub-datasets for their case.

¹ <https://github.com/maja601/EuroCrops/pull/5>

Tackling label meagre Hyperspectral Imaging (HSI) challenges with great quantities of Multispectral Imaging (MSI) labels Gao et al. (2023) developed a new method for pixel-wise HSI classification with domain generalisation, assuming different spectra of the same material were different domains, and a new loss function, specifically designed for data suffering from hyperspectral heterospectra. While the most common HSI datasets hardly offer any meaningful spatial extent, the experiments on the method were extended to the multispectral case. They used TINYEURO-CROPS and separated the labels into two classes: arable crops and permanent crops. They then defined three non-overlapping domain labels for each of these classes, simulating a comparable problem as hyperspectral heterospectra, helping the authors validate their methodology.

Multimodal, -year, and -country crop type classification Barriere, Claverie, Schneider, Lemoine and d'Andrimont (2024) fuse SITS from Sentinel-2 and Landsat 8 with parcel-wise Crop Rotation (CR) information and the local distribution of crops. In contrast to many commonly used deep learning practises that only use RS data, their approach includes information about agricultural practices that can bridge the gap when there is no data and enrich the decision process with historical knowledge of CR and crop distributions of surrounding parcels from earlier seasons. Additionally, they developed a new method to automatically find clusters in the crop labels, based on the hierarchy embedded in EURO-CROPS and the lessons learned while working on HCAT. This enabled them to benefit from the pan-European harmonisation of the data and run transfer learning zero-shot and few-shot experiments with France and the Netherlands. Figure 3.1 visualises the cross-country adaption of labels.

Assessment and validation of foundation models for RS Li et al. (2024) used TINYEURO-CROPS to validate their evaluation of AI foundation models and their baseline application of permafrost mapping. They show that if these models are fine-tuned to geospatial tasks, they achieve high accuracy and, by using data from EURO-CROPS, they can conclude that the advantages and challenges of using such models are generally applicable and not just for a single use-case. With the rise of data-hungry foundation models, large datasets like this play an integral role in their development and assessment.

MODIS NDVI yield forecast validation With a new pixel-based approach, Seguini et al. (2024) produced a European-wide winter crop map for early-season yield forecasting. EURO-CROPS's HCAT is the basis for identifying winter crops across the EU, alongside additional data collected as an extension to be used for validation.

Land cover mapping with domain adaptation, contrastive learning and feature disentanglement Dantas et al. (2024) presented a new DL method for reusing historical ground truth data of land cover maps to improve results on new predictions from SITS. They employed several recently developed methodologies, such as domain adaptation to fuse data from the previous years and the current one and a contrastive learning strategy for disentanglement, which supports feature extraction from the specific domains. The authors tested their proposed methods on two study sites, one in Burkina Faso and one in France, to ensure a large variety of landscapes. Half of the French data is sourced from EURO-CROPS, highlighting again the importance of keeping the dataset partitioned but effortlessly accessible to drive methodological developments in DL for RS.

Large scale pesticide exposure monitoring For their study on pesticide exposure risk, Galimberti et al. (2024) analysed a large variety of datasets and promoted a new model that aids in estimating the local impact of pesticide use in agricultural areas on the population. This model can be applied to whole countries and stresses the importance of pesticide monitoring on local scales for policymakers

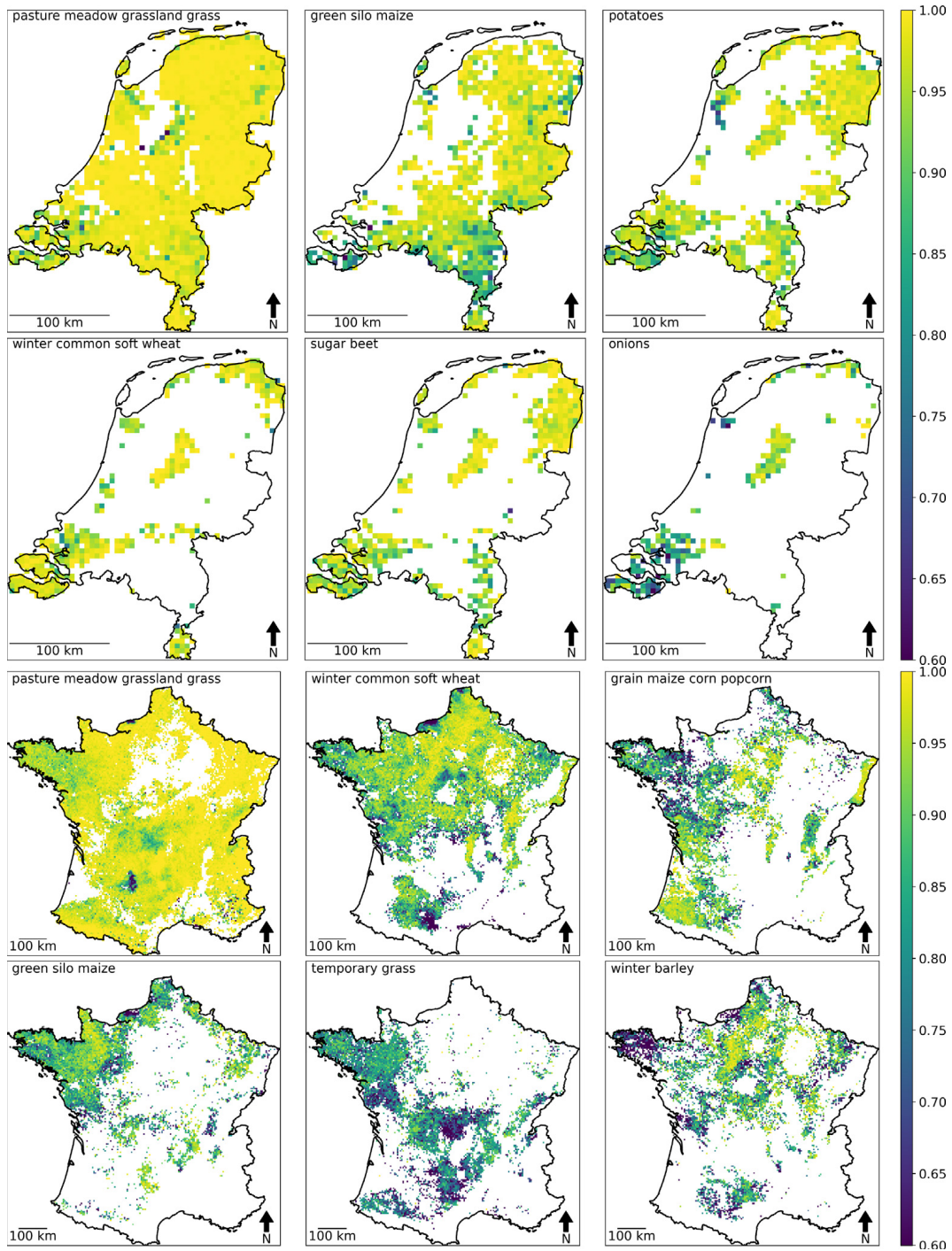


Figure 3.1: Barriere et al. (2024) use EUROCRIPS and HCAT to showcase the transnational applicability of their crop type classification method. Here, the F1-score for each crop in a 5 km grid cell is computed and visualised for the Netherlands (top) and France (bottom). The specific crop classes in the left top corner of each image are taken from the HCAT taxonomy and can be found in Figure 2.4. Source: (Barriere et al. 2024)

to ensure safety for the public. Here, EUROCROPS helped by providing a semantic consistency between crop classes and shows that HCAT is vital in assisting researchers to advance in their field and give recommendations for the future.

Building global datasets with a single European data source Parente et al. (2024) were able to use EUROCROPS as a sub-collection for a global reference grassland dataset. To use all available European datasets in one go, the authors only had to develop a mapping from HCAT to their grassland taxonomy. This way, they could assess their global grassland maps derived from EO data, high-resolution imagery and pre-existing samples and maps. Compared to the other reference datasets used, the method yields a similar performance on EUROCROPS, which indicates that the dataset is of equal quality and usability.

4 Impact of contributions and recommendations

EUROCROPS has not only been the driver of methodological developments in ML for RS but also a trailblazer for developing open pan-European datasets. The initiative has shown what is possible, and the research world has begun vocalising its need for this type of work. There will now be a more general view of the impact, less from a methodological, but from a universal perspective. Therefore, voices from research will be included that show the size of the gap that EUROCROPS partially albeit successfully filled in. In a subsequent section, the focus will be on the recommendations based on the experiences of working with pan-European data and the requirements for similar initiatives.

4.1 Impact of contributions and researchers opinions

Inadequate geographic diversity and spatial extent in crop datasets Kondmann et al. (2021) worked on a dataset themselves but realised the importance of balancing temporal and spatial resolution with the requirement to cover a large area. They put EUROCROPS forward as part of a potential solution to the geographic generalisability problem and highlight the importance of large area-covering datasets.

Information base for quantitative validation of methods As previously mentioned, the EUROCROPS initiative not only included the creation of the dataset but also a large variety of collected information, ranging from data sources, licenses and detailed information of the individual sub-datasets content. Fendrich et al. (2023), for example, use this collected knowledge to find background information about the countries' LPIS/GSA declaration style to evaluate their methods for cover crop maps in Europe. This reveals another critical impact of the project: Gathering information, formatting it, and making it easily accessible to a large group of people. This drives transnational research in a way that outdoes the effort to create such a database. The barrier that scattered information imposes on the academic community actively limits the advances that could have been made in a field, as time and funding are often finite. The mere gathering of resources is often seen as insufficient for scientific publications, inadequate without a new method introduced in combination.

Framework for challenges and opportunities of transnational datasets During the development of the dataset, it was possible to directly identify the most significant barriers of working with transnational data and institutes, as well as the varied benefits that the outcomes could hold (Schneider, Marchington et al. 2022). The six identified challenges, as well as the solution within the EUROCROPS ecosystem, were the following:

1. **Discoverability** Data scattered across Europe is hard to find, but the elaborate documentation on the EUROCROPS Wiki (See reference in Chapter 2.4.3) page holds all information about data sources together.
2. **Accessibility** Inhomogeneous legislation in MSs hinders the publication of similar datasets, but pushing the boundaries by example helped policymakers to take the need for this type of data seriously. EUROCROPS reference data being available on Zenodo (Schneider, Chan et al. 2023), but also platforms like EO-Lab¹, CODE-DE², EuroDataCube³, Source Cooperative⁴,

¹ <https://eo-lab.org/en/portfolio/?q=Training-%26-Reference-Data>

² <https://code-de.org/en/portfolio/>

³ <https://collections.eurodatacube.com/eurocrops/>

⁴ <https://beta.source.coop/repositories/cholmes/eurocrops/description/>

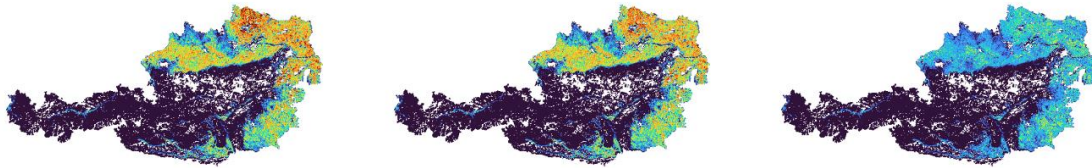


Figure 4.1: The number of different crop parcels per grid box in Austria. Red colour means more, blue to purple are the areas with less. On the left, most classes available in the original Austrian dataset are considered. The middle figure shows a truncated version where, for example, seasonal crops are not distinguished and yielding to less red areas in the northern part. The figure on the right shows the number of different crops on a low HCAT level, making it easily comparable to a dataset from a country with a coarse crop distinction in its raw data. Source (Schneider, Gackstetter et al. 2025)

ensures that researchers get easy access to the data from many venues. This was supported by the ongoing efforts for the development of high-value dataset regulations.

3. **Homogeneity** With HCAT, bridging country-dependent differences in taxonomies and creating one European foundation for cultivated crops became possible.
4. **Scalability** As HCAT goes through constant development stages, it is possible to tackle the issue of growing databases. New datasets can be easily integrated into the scheme, and additional classes can be added.
5. **Distributability** Once again, by highlighting the difficulties of obtaining theoretically publicly available data, the initiative hopes to shed light on the importance of having central European data hubs that are being used.
6. **Maintainability** Following the previous point, local and regional servers undergo a large variety of changes over time, making it difficult to track the data sources. With EUROCRIPS, the plan was to always keep the data sources and licences with the data in case changes arise, but even then, hyperlinks suddenly point to dead ends and data gets updated without notifications. This challenge is difficult to tackle from an outside perspective and requires joint efforts to ensure changes are passed to the relevant stakeholders.

Using HCAT to help inform biodiversity research Schneider, Gackstetter, Prexl, Meyer and Körner (2025) investigated the broader impact of having a harmonised pan-European dataset with all crop classes hierarchically organised. The work focused on the information included in the vector data without utilising corresponding RS imagery, assessing the insights obtained with regard to biodiversity drivers. They conducted a case study where several agricultural diversity indices for 1 km x 1 km grids were computed, the scope being countries included in EUROCRIPS and harmonised within HCAT. By leveraging the different hierarchical levels of HCAT, they were able to successfully compare countries whose original had varying levels of granularity. Figure 4.1 illustrates the amount of different crop parcels for the three different hierarchical HCAT levels, ranging from fine granularity (left) to coarse granularity (right) for Austria.

4.2 Recommendations

Data access and management in the EU is constantly undergoing changes with a clear trend towards open data and data sharing. In order to keep that development moving, this section will briefly state

the recommendations that can be taken away from a pan-European project such as EURO-CROPS. These can be seen as a secondary result of the scientific work and shall provide future initiative guidelines.

Standardisation and harmonisation

- **Language in datasets** Adopt a common language framework to keep consistency and interoperability across different regions and platforms. By using a common framework, cross-communication of data between different countries becomes far more feasible.
- **Provide taxonomy** Develop and use a standardised agricultural taxonomy to ensure comparable classification and analysis. This way, reliable and reproducible results on different scales can be achieved as the reference frame prevails.
- **EU taxonomy with mappings** In addition to a provided taxonomy, mappings to existing taxonomies enable researchers to make use of different data sources. Temporal upward and downward compatibility of data attributes can be ensured and allows the use of historical data in combination with newer ones.
- **Pan-European data harmonisation** Appreciate the importance of pan-European data harmonisation initiatives, not just in the agricultural sector. Data sharing has always benefited scientific developments but can only be fully exploited when the sub-collections are comparable.
- **Distribute mapping efforts** Provide taxonomy but collaborate with the MSs to obtain the best mapping. Harness the data sovereignty of the MSs by letting them participate in the mapping as they already have the best understanding of internal data structures.
- **Minimal solution: Centralised information hub** Promote one website that provides comprehensive information about the available data sources, licences and metadata. If reaching conformity for platforms, formats, and contents is too difficult, supporting transnational research by providing higher-order information can suffice.

Data access and sharing

- **Access to satellite data** Lower barriers to access satellite data for all stakeholders, including researchers and policymakers. By removing the exclusivity of working with EO data, interdisciplinary work and faster prototyping are possible.
- **Cloud platforms** Advocate for the use of cloud platforms for data storage and processing to enable scalable and efficient data handling. While reducing the redundancy of local data storage, sharing computing can also enable resource-conscious working.
- **Support regional LPIS and ministries** Provide better support for MSs to help them achieve better data collection and management capabilities. Clarifying goals and processes through bilateral communication enables employees of public institutions to better understand the objectives without feeling isolated.
- **Data lake for EU** Instead of complex platforms, establish a centralised data lake for the EU to aggregate and store diverse datasets. Similar to the centralised information hub, a data lake could offer a low-level solution for collecting data.
- **Rich metadata** Put stress on the utilisation of rich metadata in order to provide context and improve data usability. Specify clear examples, guidelines and reasons for contributors to understand the impact of having metadata at hand.

Data utilisation and processing

- **Provide ready-to-use data for interdisciplinary work** Encourage scientists from different fields to partake in RS research by lowering the entry barrier of working with RS data. Easily downloadable benchmark datasets encourage the development of state-of-the-art methods and invite people from other disciplines to join.
- **Publish models and pipelines** Enable MSs to make full use of satellite and agricultural data by providing the means to process them. Having platforms with trained models and data analysing procedures that support governmental work gives an incentive to further provide open data to feed into the loop.
- **Collaborate for a singular large dataset** Put effort into compiling comprehensive, high-quality datasets that support multiple use cases and researchers and policymakers instead of a lot of scattered, small ones. This way, interoperability and a larger acceptance can be achieved.

Policy and regulation

- **Communication regarding GDPR** Reach out to MSs and actively work on a mutual understanding of data privacy and the potential benefits of sharing. Laying out clear objectives that both the research community and MSs can benefit from is the basis for large-scale transnational datasets.
- **Transition times without strict regulations** Allow for flexible transition periods while encouraging data sharing and accepting intermediate results. Introduce open data step-by-step to reluctant MSs by showing them the benefits of each phase and grant a not-perfect approach.

Platform development and community engagement

- **One single platform for information exchange** Work towards establishing e.g. INSPIRE as a single, centralised platform for exchanging information, tools and recuses. By putting effort into keeping all material up-to-date and leading by example by using such a platform, stakeholders will gain trust to use it as a central information point over time.
- **Platforms and their utilisation** Investigate why existing platforms are not taken advantage of and develop strategies to make one a standard by providing examples and support. Obtaining the reason for reluctance and working actively against it can yield faster results than working out a new system.
- **Allow community participation** Let the community be part of the data collection, processing and analysis through open and transparent processes and feedback opportunities. Allow for discussion and GitHub-like issues in order to encourage community efforts and crowd intelligence in order to raise data quality and quantity.

Technical and operational recommendations

- **Ensure satellite operability** Guarantee satellite systems' continuous operability and reliability to provide data in regular, uninterrupted intervals. Researchers and policymakers must build trust in the applicability of EO data for their cause.
- **Different data formats: Standards and preferences** Investigate the benefits and drawbacks of having different data formats and establish standards to enable data sharing and integration. Provide an answer as to whether the amount of distinct geospatial data formats is necessary or if there is a possibility of achieving common ground.

- **Keep use case agnostic** Develop datasets and tools that are agnostic to specific use cases to facilitate various applications. Having a new process for every research question or operation appears feasible on small scales but not on large pan-European ones.

Research and collaboration

- **Researchers work with real-world data** Let researchers work on relevant and impactful outcomes by providing usable real-world data. Theoretical groundwork often lacks the connection to applications and can benefit from easily accessible small-scale examples that could be magnified by arbitrary factors.
- **Talk to experts but implement the minimal solutions** Consult field experts while prioritising the implementation of a minimal, widely agreed-upon solution to ensure acceptance, feasibility and progression. Avoid dead-lock situations by continuously working on compromises.
- **Establish communication between researchers and regional authorities** Let both sides see the advantages of collaboration and positive impact by fostering communication between all stakeholders. Collaboration between the public body and research ensures the best understanding of data and, therefore, outcomes.

5 Conclusion and outlook

Conclusion This work laid out the background, scientific problems and impacts related to the EUROCRIPS dataset as an advocate for open data initiatives, particularly in the RS context. The primary research question focused on the work with heterogeneous, large-scale, geospatial, administrative data collections and the pathway to developing widely accepted reference datasets for a large variety of use cases and verifying ML methodologies.

Integrating administrative data significantly contributed to the recent developments in ML for RS. Chapter 1 laid out the potential of satellite imagery for LCLU classification and highlighted the advancements made possible by using such data. Chapter 2 provided an intuition for agricultural subsidy control data, as well as data collection and harmonisation in Europe, leading to the discussion of the background, scientific question and key insights of EUROCRIPS. These include experiences in executing informed decision-making while different types of stakeholders are involved, as well as identifying the importance of flexible but extensive datasets. Methodological developments for cropland classification based on EUROCRIPS are introduced in Chapter 3, once again highlighting the now frequent and successful use of the dataset despite being published for only a few years. Chapter 4 showcases the datasets' broader impact outside of LCLU classification tasks, confirming it as a piece of pan-European, large-scale dataset research. As an answer to the question about how to approach similar initiatives in the future, this work closes with recommendations for transnational data projects, spanning seven categories: Standardisation and harmonisation; data access and sharing; data utilisation and processing; policy and regulation; platform development and community engagement; technical and operational recommendations; and research and collaboration.

This work contributes to the field by providing a comprehensive meta-analysis around the EUROCRIPS initiative, facilitating researchers and policymakers to take away insights of working on such data, without having to undergo the entire process again. Additionally, by providing clear timelines and connections between publications, it summarised the efforts that are connected to the largest pan-European crop dataset.

The recommendations and findings have broad implications for a large variety of stakeholders, spanning from applications like agricultural monitoring and method development in ML to abstractly working in transnational research. This allows them to be applied to a multitude of data initiatives and within decision-making processes.

However, the amount of suggestions gives an indication of how much work remains missing. The EUROCRIPS initiative cannot be seen as the solution to all issues arising when working on a pan-European level, but more as a starting point. Going forward, there are still large obstacles that need to be tackled, exceeding the presented work as transnational data initiatives are still in their infancy. Nevertheless, the presented work opens the stage for discussions and efforts towards harmonised pan-European datasets.

To conclude, this work demonstrates the importance of the scientific exploration conducted throughout the course of the EUROCRIPS initiative. It showcases the fundamental usage and impact of such a dataset and clearly exemplifies the need for continued development within this sector.

Outlook Although initially started as a pilot, EUROCRIPS is now an integral part of pan-European agricultural research. The support provided by the JRC of the EC is now evolving into a full cooperation between the Technical University of Munich (TUM) and the JRC. The CHEAP database developed by Claverie et al. (2024) will be combined with the insights from EUROCRIPS, alongside a potential update of HCAT. Eventually, a transition towards an operational system rather than a research initiative will be set in motion, enabling scientists in the future generations to take part in this pan-European research.

6 Summary of publications

Publications that are referenced in and contributed to the dissertation are summarised on the following pages. The chapters that contain the contribution of each publication are referenced in the "Relevant for the dissertation" paragraph of each summary.

6.1 EUROCRUPS: The Largest Harmonized Open Crop Dataset Across the European Union

Abstract EUROCRUPS contains geo-referenced polygons of agricultural croplands from 16 countries of the European Union (EU) as well as information on the respective crop species grown there. These semantic annotations are derived from self-declarations by farmers receiving subsidies under the common agricultural policy (CAP) of the European Commission (EC). Over the last 1.5 years, the individual national crop datasets have been manually collected, the crop classes have been translated into the English language and transferred into the newly developed hierarchical crop and agriculture taxonomy (HCAT). EUROCRUPS is publicly available under continuous improvement through an active user community.

Author Contributions M.S. leads the EUROCRUPS project, identified data sources, created HCAT, obtained feedback from authorities and compiled the published shapefiles. T.S. obtained and documented the individual datasets from the data sources. F.S. translated the crop classes and analysed the individual datasets. M.S., T.S. and F.S. verified the crop translations and mappings to HCAT. M.K. was involved in the design of the concept and supervised the project. All authors reviewed the manuscript.

Review type Peer review

Relevance for the dissertation This work is the complementary publication to the presented dissertation. It holds all the technical details for the development of EUROCRUPS, alongside all data sources and methodologies used to obtain the dataset. Therefore, it can be seen as the core around which this dissertation is written: Chapter 1 motivated the initiative from a broader standpoint and Chapter 2 gave a more specialised background to the type of data utilised in EUROCRUPS, followed by a full meta-analysis of the dataset, rounding off the information given in the publication. In addition to the paper, this dissertation further analyses its impact, which is then presented in Chapters 3 and 4.

Citation Maja Schneider, Tobias Schelte, Felix Schmitz and Marco Körner (2023). 'EuroCrops: The Largest Harmonized Open Crop Dataset Across the European Union'. In: *Scientific Data* 10.1, p. 612. DOI: [10.1038/s41597-023-02517-0](https://doi.org/10.1038/s41597-023-02517-0)

6.2 Harnessing Administrative Data Inventories to Create a Reliable Transnational Reference Database for Crop Type Monitoring

Abstract With leaps in machine learning techniques and their application on Earth observation challenges has unlocked unprecedented performance across the domain. While the further development of these methods was previously limited by the availability and volume of sensor data and computing resources, the lack of adequate reference data is now constituting new bottlenecks. Since creating such ground-truth information is an expensive and error-prone task, new ways must be devised to source reliable, high-quality reference data on large scales. As an example, we showcase EURO-CROPS, a reference dataset for crop type classification that aggregates and harmonizes administrative data surveyed in different countries with the goal of transnational interoperability.

Author Contributions M.S.: Writing – review and editing, Writing – original draft. M.K.: Writing – review and editing, Writing – original draft.

Review type Peer review

Relevance for the dissertation This work extensively motivates the use of agricultural administrative data for research. In Chapter 2, the background of EURO-CROPS is explained with the incentive given in this referenced paper.

Citation Maja Schneider and Marco Körner (2022). 'Harnessing Administrative Data Inventories to Create a Reliable Transnational Reference Database for Crop Type Monitoring'. In: *IGARSS 2022 - 2022 IEEE International Geoscience and Remote Sensing Symposium*, pp. 5385–5388. DOI: [10.1109/IGARSS46834.2022.9883089](https://doi.org/10.1109/IGARSS46834.2022.9883089)

6.3 [Re] Satellite Image Time Series Classification with Pixel-Set Encoders and Temporal Self-Attention

Reproducibility Summary The presented study evaluates "Satellite Image Time Series Classification with Pixel-Set Encoders and Temporal Self-Attention" by Garnot et al. (2020) within the scope of the ML Reproducibility Challenge 2020. Our work focuses on both aspects constituting the paper: the method itself and the validity of the stated results. We show that, despite some unforeseen design choices, the investigated method is coherent in itself and performs the expected way.

Scope of Reproducibility: The evaluated paper presents a method to classify crop types from multispectral satellite image time series with a newly developed pixel-set encoder and an adaption of the Transformer, called temporal attention encoder.

Methodology: In order to assess both the architecture and the performance of the approach, we first attempted to implement the method from scratch, followed by a study of the authors openly provided code. Additionally, we also compiled an alternative dataset similar to the one presented in the paper and evaluated the methodology on it.

Results: During the study, we were not able to reproduce the method due to a conceptual misinterpretation of ours regarding the authors adaption of the Transformer. However, the publicly available implementation helped us answering our questions and proved its validity during our experiments on different datasets. Additionally, we compared the papers temporal attention encoder to our adaption of it, which we came across while we were trying to reimplement and grasp the authors ideas.

What was easy: Running the provided code and obtaining the presented dataset turned out to be easily possible. Even adapting the method to our own ideas did not cause issues, due to a well documented and clear implementation.

What was difficult: Reimplementing the approach from scratch turned out to be harder than expected, especially because we had a certain type of architecture in mind that did not fit the dimensions of the layers mentioned in the paper. Furthermore, knowing how the dataset was exactly assembled would have been beneficial for us, as we tried to retrace these steps, and therefore would have made the results on our dataset easier to compare to the ones from the paper.

Communication with original authors: While working on the challenge, we stood in E-mail contact with the first and second author, had two online meetings and got feedback to our implementation on GITHUB. Additionally, one of the authors of the Transformer paper provided us with further answers regarding their models architecture.

Author Contributions M.S.: Writing – review and editing, Writing – original draft. M.K.: Writing – review and editing.

Review type Peer review

Relevance for the dissertation This publication first employed the TINYEUROCROPS dataset for verifying methodological assumptions. While Garnot et al. (2020) trained and tested their methods in a relatively small regional scale, this work showed how the technique performed with different test data origins. In Chapter 3, the work is used as a first demonstrator to the applicability of the data.

Citation Maja Schneider and Marco Körner (2021 a). '[Re] Satellite Image Time Series Classification with Pixel-Set Encoders and Temporal Self-Attention'. In: *ReScience C* 7.2, #19. DOI: [10.5281/zenodo.4835356](https://doi.org/10.5281/zenodo.4835356)

6.4 Challenges and Opportunities of Large Transnational Datasets: A Case Study on European Administrative Crop Data

Abstract Expansive, informative datasets are vital in providing foundations and possibilities for scientific research and development across many fields of study. Assembly of grand datasets, however, frequently poses difficulty for the author and stakeholders alike, with a variety of considerations required throughout the collaboration efforts and development lifecycle. In this work, we discuss and analyse the challenges and opportunities we faced throughout the creation of a transnational, European agricultural dataset containing reference labels of cultivated crops. Together, this forms a succinct framework of important elements one should consider when forging a dataset of their own.

Author Contributions M.S.: Writing – review and editing, Writing – original draft. C.M.: Writing – review and editing, Writing – original draft. M.K.: Writing – review and editing.

Review type Peer review

Relevance for the dissertation The challenges and opportunities identified in this paper are a direct result of working on large transnational datasets. In this dissertation, they are highlighted in Section 4.1, showing the interdisciplinary impact and contribution of EURO CROPS and HCAT. The scientific contribution of the work has gone beyond the use case of agricultural datasets, influencing international research collaborations.

Citation Maja Schneider, Christian Marchington and Marco Körner (2022). ‘Challenges and Opportunities of Large Transnational Datasets: A Case Study on European Administrative Crop Data’. In: *Workshop on Broadening Research Collaborations in ML (NeurIPS 2022)*. DOI: [10.48550/arXiv.2210.07178](https://doi.org/10.48550/arXiv.2210.07178)

6.5 Advancing Transnational Assessments of Biodiversity Drivers in European Agriculture with an Updated Hierarchical Crop and Agriculture Taxonomy (HCAT)

Abstract Modern agriculture plays a significant role in driving the decline of global biodiversity. The homogenization of landscapes, the reduction of natural habitats, and the intense use of pesticides are substantial factors for natural species populations to shrink or even disappear. However, despite significant advances in research, still today, the impacts of cropping systems on biodiversity are challenging to quantify. One primary reason for this is the lack of available agricultural data. The data from the Integrated Administration and Control System (IACS) of the European Union's (EU) Common Agricultural Policy (CAP) give new potential to improve the basis of information for agroecological research in Europe. Within the framework of the CAP, European farmers are required to declare their cropping arrangements to official authorities to receive corresponding subsidies in exchange. The nationally applied crop taxonomies are, however, not harmonized across Europe, which hinders transnational analyses of agriculture and its environmental impacts. To overcome this barrier, we developed a Hierarchical Crop and Agriculture Taxonomy (HCAT) to harmonize administrative, agricultural data from 16 EU member states. With the release of our upgraded second version of HCAT, we demonstrate how a harmonized CAP data set can aid in identifying drivers of biodiversity in agricultural landscapes at both national and international scales.

Author Contributions M.S. leads the EURO-CROPS project, where she collected the data and developed the taxonomy. D.G., M.S. and S.M. developed the concept for introducing the taxonomy into the agricultural and ecological context, including the discussion on HCATv2's implications for practical application. D.G. and M.S. developed the experimental design. M.S. and D.G. prepared the visualizations and tables. Together with D.G., J.P. developed the software necessary for the case study. All authors contributed to writing the paper.

Review type Peer review

Relevance for the dissertation This paper highlights the wide range of use cases of transnational datasets. This work gained scientific insights by analysing the vector data and showing the importance of hierarchical harmonised datasets across the EU, as mentioned in Section 4.1.

Citation Maja Schneider, David Gackstetter, Jonathan Prexl, Sebastian T. Meyer and Marco Körner (2025). 'Advancing Transnational Assessments of Biodiversity Drivers in European Agriculture with an Updated Hierarchical Crop and Agriculture Taxonomy (HCAT)'. in: *npj Sustainable Agriculture* 3.1, pp. 1–10. DOI: [10.1038/s44264-024-00037-x](https://doi.org/10.1038/s44264-024-00037-x)

List of abbreviations

- AI** Artificial Intelligence
- CAP** Common Agricultural Policy
- CNN** Convolutional Neural Network
- CR** Crop Rotation
- DL** Deep Learning
- EC** European Commission
- EEA** European Environment Agency
- EO** Earth Observation
- EU** European Union
- GDPR** General Data Protection Regulation
- GEDI** Global Ecosystem Dynamics Investigation
- GIS** Geographic Information System
- GRU** Gated Recurrent Unit
- GSA** Geospatial Application
- HCAT** Hierarchical Crop and Agriculture Taxonomy
- HSI** Hyperspectral Imaging
- IACS** Integrated Administration and Control System
- INSPIRE** Infrastructure for Spatial Information in Europe
- JRC** Joint Research Centre
- LC** Land Cover
- LCC** Land Cover Classification
- LCLU** Land Cover/Land Use
- LPIS** Land Parcel Identification System
- LSTM** Long Short-Term Memory
- LU** Land Use
- ML** Machine Learning
- MLE** Maximum Likelihood Estimation
- MLP** Multilayer Perceptron

MS Member State
MSI Multispectral Imaging
NDVI Normalized Difference Vegetation Index
NN Neural Network
PCA Principal Component Analysis
RF Random Forest
RNN Recurrent Neural Network
RS Remote Sensing
SAR Synthetic-Aperture Radar
SITS Satellite Image Time Series
SVM Support Vector Machine
TOA Top Of Atmosphere

Personal Publications

All personal publications, including the ones presented earlier, are listed here. These include works, that were not explicitly mentioned in this dissertation, but contain insights gained while working on EURO-CROPS.

2025

- Maja Schneider, David Gackstetter, Jonathan Prexl, Sebastian T. Meyer and Marco Körner (2025). 'Advancing Transnational Assessments of Biodiversity Drivers in European Agriculture with an Updated Hierarchical Crop and Agriculture Taxonomy (HCAT)'. in: *npj Sustainable Agriculture* 3.1, pp. 1–10. DOI: [10.1038/s44264-024-00037-x](https://doi.org/10.1038/s44264-024-00037-x)

2024

- Valentin Barriere, Martin Claverie, Maja Schneider, Guido Lemoine and Raphaël d'Andrimont (2024). 'Boosting Crop Classification by Hierarchically Fusing Satellite, Rotational, and Contextual Data'. In: *Remote Sensing of Environment* 305, p. 114110. DOI: [10.1016/j.rse.2024.114110](https://doi.org/10.1016/j.rse.2024.114110)

2023

- Ayshah Chan, Maja Schneider and Marco Körner (2023). 'XAI for Early Crop Classification'. In: *IGARSS 2023 - 2023 IEEE International Geoscience and Remote Sensing Symposium*, pp. 2657–2660. DOI: [10.1109/IGARSS52108.2023.10281498](https://doi.org/10.1109/IGARSS52108.2023.10281498)
- Maja Schneider, Tobias Schelte, Felix Schmitz and Marco Körner (2023). 'EuroCrops: The Largest Harmonized Open Crop Dataset Across the European Union'. In: *Scientific Data* 10.1, p. 612. DOI: [10.1038/s41597-023-02517-0](https://doi.org/10.1038/s41597-023-02517-0)
- Bénédicte Bucher, Marcin Grudzień, Nathalie Delattre, Jordi Escriu Paradell, Erwin Folmer, Antonin Garrone, Antje Kügeler, Ángel Lopez, Ed Parsons, Andrea Perego, Jiri Pilar, Jari Reini, Hannes I. Reuter, Jill Saligoe-Simmel, Maja Schneider and Jeroen Ticheler (2023). 'Geodata Discoverability'. In: *Joint Workshop of EuroGeographics and EuroSDR*

2022

- Maja Schneider, Christian Marchington and Marco Körner (2022). 'Challenges and Opportunities of Large Transnational Datasets: A Case Study on European Administrative Crop Data'. In: *Workshop on Broadening Research Collaborations in ML (NeurIPS 2022)*. DOI: [10.48550/arXiv.2210.07178](https://doi.org/10.48550/arXiv.2210.07178)
- Maja Schneider and Marco Körner (2022). 'Harnessing Administrative Data Inventories to Create a Reliable Transnational Reference Database for Crop Type Monitoring'. In: *IGARSS 2022 - 2022 IEEE International Geoscience and Remote Sensing Symposium*, pp. 5385–5388. DOI: [10.1109/IGARSS46834.2022.9883089](https://doi.org/10.1109/IGARSS46834.2022.9883089)

2021

- Maja Schneider, Amelie Broszeit and Marco Körner (2021). 'EuroCrops: A Pan-European Dataset for Time Series Crop Type Classification'. In: *Proceedings of the Conference on Big Data from Space (BiDS)*. DOI: [10.2760/125905](https://doi.org/10.2760/125905)
- Maja Schneider and Marco Körner (2021a). '[Re] Satellite Image Time Series Classification with Pixel-Set Encoders and Temporal Self-Attention'. In: *ReScience C* 7.2, #19. DOI: [10.5281/zenodo.4835356](https://doi.org/10.5281/zenodo.4835356)

Bibliography

- Arnold, Stephan, Barbara Kosztra, Gebhard Banko, Geoff Smith, Gerard Hazeu, Michael Bock and Nuria Valcarcel Sanz (2013). 'The EAGLE Concept – A Vision of a Future European Land Monitoring Framework'. In: *Proceedings of the 33rd EARSeL Symposium towards Horizon, Matera, Italy*.
- Aroma, R. Jenice and Kumudha Raimond (2016). 'An Overview of Technological Revolution in Satellite Image Analysis'. In: *Journal of Engineering Science and Technology Review* 9.4, pp. 1–5. DOI: [10.25103/jestr.094.01](https://doi.org/10.25103/jestr.094.01).
- Aszkowski, Przemysław and Marek Kraft (2023). 'Challenges of Crop Classification from Satellite Imagery with Eurocrops Dataset'. In: *Progress in Polish Artificial Intelligence Research* 4. DOI: [10.34658/9788366741928.2](https://doi.org/10.34658/9788366741928.2).
- Baiamonte, Giuseppe, Gilbert Voican and Philippe Loudjani (2023). 'Getting the Most of Land Parcel Identification Systems (LPIS) and GeoSpatial Aid Application (GSAA) Datasets'. European Commission, Ispra, 2023, JRC133145.
- Barrett, Samuel and Ana Toro (2024). 'Data-Centric Solutions to Applied Crop Type Classification at Scale: Towards a Globally Applicable Many-Crop Model'. Presented at EO for Agriculture under Pressure, ESA-ESRIN Frascati, Italy.
- Barriere, Valentin, Martin Claverie, Maja Schneider, Guido Lemoine and Raphaël d'Andrimont (2024). 'Boosting Crop Classification by Hierarchically Fusing Satellite, Rotational, and Contextual Data'. In: *Remote Sensing of Environment* 305, p. 114110. DOI: [10.1016/j.rse.2024.114110](https://doi.org/10.1016/j.rse.2024.114110).
- Bucher, Bénédicte, Marcin Grudzień, Nathalie Delattre, Jordi Escriu Paradell, Erwin Folmer, Antonin Garrone, Antje Kügeler, Ángel Lopez, Ed Parsons, Andrea Perego, Jiri Pilar, Jari Reini, Hannes I. Reuter, Jill Saligoe-Simmel, Maja Schneider et al. (2023). 'Geodata Discoverability'. In: *Joint Workshop of EuroGeographics and EuroSDR*.
- Chan, Ayshah, Maja Schneider and Marco Körner (2023). 'XAI for Early Crop Classification'. In: *IGARSS 2023 - 2023 IEEE International Geoscience and Remote Sensing Symposium*, pp. 2657–2660. DOI: [10.1109/IGARSS52108.2023.10281498](https://doi.org/10.1109/IGARSS52108.2023.10281498).
- Claverie, Martin, Valentin Barriere, Raphaël d'Andrimont, Renate Koble and Marijn Van der Velde (2024). 'In-season Crop Type Mapping: An accuracy evaluation at European scale using the CHEAP Database'. Presented at EO for Agriculture under Pressure, ESA-ESRIN Frascati, Italy.
- Connelly, Roxanne, Christopher J. Playford, Vernon Gayle and Chris Dibben (2016). 'The role of administrative data in the big data revolution in social science research'. In: *Social Science Research. Special issue on Big Data in the Social Sciences* 59, pp. 1–12. DOI: [10.1016/j.ssresearch.2016.04.015](https://doi.org/10.1016/j.ssresearch.2016.04.015).
- D'Andrimont, Raphaël, Momchil Yordanov, Laura Martinez Sanchez, Peter Haub, Olivier Buck, Carsten Haub, Beatrice Eiselt and Marijn Van der Velde (2022). 'LUCAS Cover 2006-2018'. JRC Publications Repository.
- Dantas, Cassio F., Raffaele Gaetano, Claudia Paris and Dino Ienco (2024). 'Reuse Out-of-Year Data to Enhance Land Cover Mapping via Feature Disentanglement and Contrastive Learning'. DOI: [10.48550/arXiv.2404.11114](https://doi.org/10.48550/arXiv.2404.11114). preprint.

- Di Tommaso, Stefania, Sherrie Wang, Vivek Vajipey, Noel Gorelick, Rob Strey and David B. Lobell (2023). 'Annual Field-Scale Maps of Tall and Short Crops at the Global Scale Using GEDI and Sentinel-2'. In: *Remote Sensing* 15.17 (17), p. 4123. DOI: [10.3390/rs15174123](https://doi.org/10.3390/rs15174123).
- Doiron, Dany, Paul Burton, Yannick Marcon, Amadou Gaye, Bruce H.R. Wolffenbuttel, Markus Perola, Ronald P. Stolk, Luisa Foco, Cosetta Minelli, Melanie Waldenberger, Rolf Holle, Kirsti Kvaløy, Hans L. Hillege, Anne-Marie Tassé, Vincent Ferretti et al. (2013). 'Data Harmonization and Federated Analysis of Population-Based Studies: The BioSHaRE Project'. In: *Emerging Themes in Epidemiology* 10.1. DOI: [10.1186/1742-7622-10-12](https://doi.org/10.1186/1742-7622-10-12).
- Esteves, Antonio and Nuno Valente (2024). 'Automatic Generation of a Portuguese Land Cover Map with Machine Learning'. In: *Intelligent Systems and Applications*, pp. 36–58. DOI: [10.1007/978-3-031-47721-8_3](https://doi.org/10.1007/978-3-031-47721-8_3).
- European Commission (2022). 'Commission Implementing Regulation (EU) 2023/138 of 21 December 2022 Laying down a List of Specific High-Value Datasets and the Arrangements for Their Publication and Re-Use (Text with EEA Relevance)'.
- Fendrich, Arthur Nicolaus, Francis Matthews, Elise Van Eynde, Marco Carozzi, Zheyuan Li, Raphaël d'Andrimont, Emanuele Lugato, Philippe Martin, Philippe Ciais and Panos Panagos (2023). 'From Regional to Parcel Scale: A High-Resolution Map of Cover Crops across Europe Combining Satellite Data with Statistical Surveys'. In: *Science of The Total Environment* 873, p. 162300. DOI: [10.1016/j.scitotenv.2023.162300](https://doi.org/10.1016/j.scitotenv.2023.162300).
- Foerster, Saskia, Klaus Kaden, Michael Foerster and Sibylle Itzerott (2012). 'Crop Type Mapping Using Spectral–Temporal Profiles and Phenological Information'. In: *Computers and Electronics in Agriculture* 89, pp. 30–40. DOI: [10.1016/j.compag.2012.07.015](https://doi.org/10.1016/j.compag.2012.07.015).
- Galimberti, Francesco, Stephanie Bopp, Alessandro Carletti, Rui Catarino, Martin Claverie, Pietro Florio, Alessio Ippolito, Arwyn Jones, Flavio Marchetto, Michael Olvedy, Alberto Pistocchi, Astrid Verhegghen, Marijn Van Der Velde, Diana Vieira and Raphaël d'Andrimont (2024). 'From Parcels to People: Development of a Spatially Explicit Risk Indicator to Monitor Residential Pesticide Exposure in Agricultural Areas'. DOI: [10.48550/arXiv.2402.10990](https://doi.org/10.48550/arXiv.2402.10990). preprint.
- Gao, Zhe, Bin Pan, Xia Xu, Tao Li and Zhenwei Shi (2023). 'LiCa: Label-Indicate-Conditional-Alignment Domain Generalization for Pixel-Wise Hyperspectral Imagery Classification'. In: *IEEE Transactions on Geoscience and Remote Sensing* 61, pp. 1–11. DOI: [10.1109/TGRS.2023.3300688](https://doi.org/10.1109/TGRS.2023.3300688).
- Garnot, Vivien Sainte Fare, Loic Landrieu, Sebastien Giordano and Nesrine Chehata (2019). 'Time-Space Tradeoff in Deep Learning Models for Crop Classification on Satellite Multi-Spectral Image Time Series'. In: *IGARSS 2019 - 2019 IEEE International Geoscience and Remote Sensing Symposium*, pp. 6247–6250. DOI: [10.1109/IGARSS.2019.8900517](https://doi.org/10.1109/IGARSS.2019.8900517).
- (2020). 'Satellite Image Time Series Classification With Pixel-Set Encoders and Temporal Self-Attention'. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12325–12334.
- Goerge, Robert M. (2018). 'Barriers to Accessing State Data and Approaches to Addressing Them'. In: *The ANNALS of the American Academy of Political and Social Science* 675.1, pp. 122–137. DOI: [10.1177/0002716217741257](https://doi.org/10.1177/0002716217741257).
- Gu, Wei, Samiul Hasan, Philippe Rocca-Serra and Venkata P. Satagopam (2021). 'Road to Effective Data Curation for Translational Research'. In: *Drug Discovery Today* 26.3, pp. 626–630. DOI: [10.1016/j.drudis.2020.12.007](https://doi.org/10.1016/j.drudis.2020.12.007).

- Gurugubelli, Venkata Sukumar, Hua Fang, James M. Shikany, Salvador V. Balkus, Joshua Rumbut, Hieu Ngo, Honggang Wang, Jeroan J. Allison and Lyn M. Steffen (2022). 'A Review of Harmonization Methods for Studying Dietary Patterns'. In: *Smart Health* 23. DOI: [10.1016/j.smhl.2021.100263](https://doi.org/10.1016/j.smhl.2021.100263).
- Hand, David J., Penny Babb, Li-Chun Zhang, Paul Allin, Anders Wallgren, Britt Wallgren, Gordon Blunt, Andrew Garrett, Fionn Murtagh, Peter W. F. Smith, Duncan Elliott, Guy Nason, Ben Powell, Jamie C. Moore, Gabriele B. Durrant et al. (2018). 'Statistical challenges of administrative and transaction data'. In: *Journal of the Royal Statistical Society. Series A (Statistics in Society)* 181.3, pp. 555–605. DOI: [10.1111/rssa.12315](https://doi.org/10.1111/rssa.12315).
- Ibrahim, Gaylan Rasul Fage, Azad Rasul and Haidi Abdullah (2023). 'Improving Crop Classification Accuracy with Integrated Sentinel-1 and Sentinel-2 Data: A Case Study of Barley and Wheat'. In: *Journal of Geovisualization and Spatial Analysis* 7.2. DOI: [10.1007/s41651-023-00152-2](https://doi.org/10.1007/s41651-023-00152-2).
- Inglada, Jordi, Marcela Arias, Benjamin Tardy, Olivier Hagolle, Silvia Valero, David Morin, Gérard Dedieu, Guadalupe Sepulcre, Sophie Bontemps, Pierre Defourny and Benjamin Koetz (2015). 'Assessment of an Operational System for Crop Type Map Production Using High Temporal and Spatial Resolution Satellite Optical Imagery'. In: *Remote Sensing* 7.9 (9), pp. 12356–12379. DOI: [10.3390/rs70912356](https://doi.org/10.3390/rs70912356).
- Kang, Jinzhong, Hongyan Zhang, Honghai Yang and Liangpei Zhang (2018). 'Support Vector Machine Classification of Crop Lands Using Sentinel-2 Imagery'. In: *2018 7th International Conference on Agro-geoinformatics (Agro-geoinformatics)*, pp. 1–6. DOI: [10.1109/Agro-Geoinformatics.2018.8476101](https://doi.org/10.1109/Agro-Geoinformatics.2018.8476101).
- Kondmann, Lukas, Aysim Toker, Marc Russwurm, Andres Camero Unzueta, Devis Peressuti, Grega Milcinski, Nicolas Longépé, Pierre-Philippe Mathieu, Timothy Davis, Giovanni Marchisio, Laura Leal-Taixé and Xiao Xiang Zhu (2021). 'DENETHOR: The DynamicEarthNET Dataset for Harmonized, Inter-Operable, Analysis-Ready, Daily Crop Monitoring from Space'. In: *35th Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, pp. 1–13.
- Kussul, Nataliia, Mykola Lavreniuk, Sergii Skakun and Andrii Shelestov (2017). 'Deep Learning Classification of Land Cover and Crop Types Using Remote Sensing Data'. In: *IEEE Geoscience and Remote Sensing Letters* 14.5, pp. 778–782. DOI: [10.1109/LGRS.2017.2681128](https://doi.org/10.1109/LGRS.2017.2681128).
- Li, Wenwen, Chia-Yu Hsu, Sizhe Wang, Yezhou Yang, Hyunho Lee, Anna Liljedahl, Chandni Witharana, Yili Yang, Brendan M. Rogers, Samantha T. Arundel, Matthew B. Jones, Kenton McHenry and Patricia Solis (2024). 'Segment Anything Model Can Not Segment Anything: Assessing AI Foundation Model's Generalizability in Permafrost Mapping'. In: *Remote Sensing* 16.5 (5), p. 797. DOI: [10.3390/rs16050797](https://doi.org/10.3390/rs16050797).
- Martirano, Giacomo and Katalin Toth (2023). 'Technical Guidelines on IACS Spatial Data Sharing - Part 2 – Interoperability'. In: *Publications Office of the European Union*. DOI: [10.2760/646422](https://doi.org/10.2760/646422).
- Mazzia, Vittorio, Aleem Khaliq and Marcello Chiaberge (2020). 'Improvement in Land Cover and Crop Classification Based on Temporal Features Learning from Sentinel-2 Data Using Recurrent-Convolutional Neural Network (R-CNN)'. In: *Applied Sciences (Switzerland)* 10.1. DOI: [10.3390/app10010238](https://doi.org/10.3390/app10010238).
- Nordbotten, Svein (2010). 'The use of administrative data in official statistics - past, present and future: with special reference to the Nordic countries'. In: *Official Statistics – Methodology and Applications in Honour of Daniel Thorburn*, pp. 205–223.

- Nyborg, Joachim, Charlotte Pelletier and Ira Assent (2022). 'Generalized Classification of Satellite Image Time Series with Thermal Positional Encoding'. In: *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1391–1401. DOI: [10.1109/CVPRW56347.2022.00145](https://doi.org/10.1109/CVPRW56347.2022.00145).
- Papadopoulou, Eleni, Giorgos Mallinis, Sofia Siachalou, Nikos Koutsias, Athanasios C. Thanopoulos and Georgios Tsaklidis (2023). 'Agricultural Land Cover Mapping through Two Deep Learning Models in the Framework of EU's CAP Activities Using Sentinel-2 Multitemporal Imagery'. In: *Remote Sensing* 15.19. DOI: [10.3390/rs15194657](https://doi.org/10.3390/rs15194657).
- Parente, Leandro, Lindsey Sloat, Vinicius Mesquita, Davide Consoli, Radost Stanimirova, Tomislav Hengl, Carmelo Bonannella, Nathália Teles, Ichsani Wheeler, Steffen Ehrmann, Maria Hunter, Laerte Ferreira, Ana Paula Mattos, Bernard Oliveira, Carsten Meyer et al. (2024). 'Mapping Global Grassland Dynamics 2000—2022 at 30m Spatial Resolution Using Spatiotemporal Machine Learning'. DOI: [10.21203/rs.3.rs-4514820/v1](https://doi.org/10.21203/rs.3.rs-4514820/v1). preprint.
- Pelletier, Charlotte, Geoffrey I. Webb and François Petitjean (2019). 'Temporal Convolutional Neural Network for the Classification of Satellite Image Time Series'. In: *Remote Sensing* 11.5 (5), p. 523. DOI: [10.3390/rs11050523](https://doi.org/10.3390/rs11050523).
- Phiri, Darius and Justin Morgenroth (2017). 'Developments in Landsat Land Cover Classification Methods: A Review'. In: *Remote Sensing* 9.9. DOI: [10.3390/rs9090967](https://doi.org/10.3390/rs9090967).
- Phiri, Darius, Matamy Simwanda, Serajis Salekin, Vincent R. Nyirenda, Yuji Murayama and Manjula Ranagalage (2020). 'Sentinel-2 Data for Land Cover/Use Mapping: A Review'. In: *Remote Sensing* 12.14. DOI: [10.3390/rs12142291](https://doi.org/10.3390/rs12142291).
- Rawat, Jiwan Singh, Vivekanand Biswas and Manish Kumar (2013). 'Changes in Land Use/Cover Using Geospatial Techniques: A Case Study of Ramnagar Town Area, District Nainital, Uttarakhand, India'. In: *The Egyptian Journal of Remote Sensing and Space Science* 16.1, pp. 111–117. DOI: [10.1016/j.ejrs.2013.04.002](https://doi.org/10.1016/j.ejrs.2013.04.002).
- Rußwurm, Marc and Marco Körner (2017). 'Multi-Temporal Land Cover Classification with Long Short-term Memory Neural Networks'. In: *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences XLII-1-W1*, pp. 551–558. DOI: [10.5194/isprs-archives-XLII-1-W1-551-2017](https://doi.org/10.5194/isprs-archives-XLII-1-W1-551-2017).
- (2018). 'Multi-Temporal Land Cover Classification with Sequential Recurrent Encoders'. In: *ISPRS International Journal of Geo-Information* 7.4 (4), p. 129. DOI: [10.3390/ijgi7040129](https://doi.org/10.3390/ijgi7040129).
- (2020). 'Self-Attention for Raw Optical Satellite Time Series Classification'. In: *ISPRS Journal of Photogrammetry and Remote Sensing* 169, pp. 421–435. DOI: [10.1016/j.isprsjprs.2020.06.006](https://doi.org/10.1016/j.isprsjprs.2020.06.006).
- Rußwurm, Marc, Charlotte Pelletier, Maximilian Zollner, Sébastien Lefèvre and Marco Körner (2020). 'BreizhCrops: A Time Series Dataset for Crop Type Mapping'. In: *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences ISPRS (2020)*. DOI: [10.5194/isprs-archives-XLIII-B2-2020-1545-2020](https://doi.org/10.5194/isprs-archives-XLIII-B2-2020-1545-2020).
- Rustowicz, Rose, Robin Cheong, Lijing Wang, Stefano Ermon, Marshall Burke and David Lobell (2019). 'Semantic Segmentation of Crop Type in Africa: A Novel Dataset and Analysis of Deep Learning Methods'. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 75–82.
- Schmitt, Michael, Lloyd Haydn Hughes, Chunping Qiu and Xiao Xiang Zhu (2019). 'SEN12MS - A Curated Dataset of Georeferenced Multi-Spectral Sentinel-1/2 Imagery for Deep Learning and Data Fu-

- sion'. In: *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences* IV-2-W7, pp. 153–160. DOI: [10.5194/isprs-annals-IV-2-W7-153-2019](https://doi.org/10.5194/isprs-annals-IV-2-W7-153-2019).
- Schmitt, Michael, Jonathan Prexl, Patrick Ebel, Lukas Liebel and Xiao Xiang Zhu (2020). 'Weakly Supervised Semantic Segmentation of Satellite Images for Land Cover Mapping-Challenges and Opportunities'. In: *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*. Vol. 5, 3, pp. 795–802. DOI: [10.5194/isprs-Annals-V-3-2020-795-2020](https://doi.org/10.5194/isprs-Annals-V-3-2020-795-2020).
- Schneider, Maja, Amelie Broszeit and Marco Körner (2021). 'EuroCrops: A Pan-European Dataset for Time Series Crop Type Classification'. In: *Proceedings of the Conference on Big Data from Space (BiDS)*. DOI: [10.2760/125905](https://doi.org/10.2760/125905).
- Schneider, Maja, Ayshah Chan and Marco Körner (2023). 'EuroCrops'. Zenodo. DOI: [10.5281/zenodo.6866846](https://doi.org/10.5281/zenodo.6866846).
- Schneider, Maja, David Gackstetter, Jonathan Prexl, Sebastian T. Meyer and Marco Körner (2025). 'Advancing Transnational Assessments of Biodiversity Drivers in European Agriculture with an Updated Hierarchical Crop and Agriculture Taxonomy (HCAT)'. In: *npj Sustainable Agriculture* 3.1, pp. 1–10. DOI: [10.1038/s44264-024-00037-x](https://doi.org/10.1038/s44264-024-00037-x).
- Schneider, Maja and Marco Körner (2021a). '[Re] Satellite Image Time Series Classification with Pixel-Set Encoders and Temporal Self-Attention'. In: *ReScience C* 7.2, #19. DOI: [10.5281/zenodo.4835356](https://doi.org/10.5281/zenodo.4835356).
- (2021b). 'TinyEuroCrops'. Technical University of Munich. DOI: [10.14459/2021MP1615987](https://doi.org/10.14459/2021MP1615987).
- (2022). 'Harnessing Administrative Data Inventories to Create a Reliable Transnational Reference Database for Crop Type Monitoring'. In: *IGARSS 2022 - 2022 IEEE International Geoscience and Remote Sensing Symposium*, pp. 5385–5388. DOI: [10.1109/IGARSS46834.2022.9883089](https://doi.org/10.1109/IGARSS46834.2022.9883089).
- Schneider, Maja, Christian Marchington and Marco Körner (2022). 'Challenges and Opportunities of Large Transnational Datasets: A Case Study on European Administrative Crop Data'. In: *Workshop on Broadening Research Collaborations in ML (NeurIPS 2022)*. DOI: [10.48550/arXiv.2210.07178](https://doi.org/10.48550/arXiv.2210.07178).
- Schneider, Maja, Tobias Schelte, Felix Schmitz and Marco Körner (2023). 'EuroCrops: The Largest Harmonized Open Crop Dataset Across the European Union'. In: *Scientific Data* 10.1, p. 612. DOI: [10.1038/s41597-023-02517-0](https://doi.org/10.1038/s41597-023-02517-0).
- Seguini, Lorenzo, Anton Vrieling, Michele Meroni and Andrew Nelson (2024). 'Annual Winter Crop Distribution from MODIS NDVI Timeseries to Improve Yield Forecasts for Europe'. In: *International Journal of Applied Earth Observation and Geoinformation* 130, p. 103898. DOI: [10.1016/j.jag.2024.103898](https://doi.org/10.1016/j.jag.2024.103898).
- Stoter, Arjan J.R., Bauke Rietveld, Vincent Jansen and Harrie J.M. Bastiaansen (2022). 'Harmonization Profiles for Trusted Data Sharing Between Data Spaces: Striking the Balance between Functionality and Complexity'. In: *CEUR Workshop Proceedings*. Vol. 3214.
- Toth, Katalin and Pavel Milenov (2020). 'Technical guidelines on IACS spatial data sharing. Part 1 - Data discovery'. In: *Publications Office of the European Union*. DOI: [10.2760/180713](https://doi.org/10.2760/180713).
- Turkoglu, Mehmet Ozgur, Stefano D'Aronco, Gregor Perich, Frank Liebisch, Constantin Streit, Konrad Schindler and Jan Dirk Wegner (2021). 'Crop Mapping from Image Time Series: Deep Learning with Multi-Scale Label Hierarchies'. In: *Remote Sensing of Environment* 264, p. 112603. DOI: [10.1016/j.rse.2021.112603](https://doi.org/10.1016/j.rse.2021.112603).

- Unim, Brigid, Elsi Haverinen, Eugenio Mattei, Flavia Carle, Andrea Faragalli, Rosaria Gesuita, Martin Thissen, Linda Abboud, Tiziana Grisetti, Petronille Bogaert and Luigi Palmieri (2022). 'Mapping European Research Networks Providing Health Data: Results from the InfAct Joint Action on Health Information'. In: *Archives of Public Health* 80.1. DOI: [10.1186/s13690-021-00766-2](https://doi.org/10.1186/s13690-021-00766-2).
- Van der Velde, Marijn (2021). 'Exploring potential of LPIS/GSAA data for JRC-MARS crop mapping, monitoring, and forecasting'. JRC note.
- Van Loenen, Bastiaan, Stefan Kulk and Hendrik Ploeger (2016). 'Data Protection Legislation: A Very Hungry Caterpillar. The Case of Mapping Data in the European Union'. In: *Government Information Quarterly* 33.2, pp. 338–345. DOI: [10.1016/j.giq.2016.04.002](https://doi.org/10.1016/j.giq.2016.04.002).
- Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser and Illia Polosukhin (2017). 'Attention Is All You Need'. In: *Advances in Neural Information Processing Systems*. Vol. 30. DOI: [10.48550/arXiv.1706.03762](https://doi.org/10.48550/arXiv.1706.03762).
- Vuolo, Francesco, Martin Neuwirth, Markus Immitzer, Clement Atzberger and Wai-Tim Ng (2018). 'How Much Does Multi-Temporal Sentinel-2 Data Improve Crop Type Classification?' In: *International Journal of Applied Earth Observation and Geoinformation* 72, pp. 122–130. DOI: [10.1016/j.jag.2018.06.007](https://doi.org/10.1016/j.jag.2018.06.007).
- Wardlow, Brian D. and Stephen L. Egbert (2008). 'Large-Area Crop Mapping Using Time-Series MODIS 250 m NDVI Data: An Assessment for the U.S. Central Great Plains'. In: *Remote Sensing of Environment* 112.3, pp. 1096–1116. DOI: [10.1016/j.rse.2007.07.019](https://doi.org/10.1016/j.rse.2007.07.019).
- Yampolskaya, Svetlana (2017). 'Research at Work: Administrative Data and Behavioral Sciences Research'. en. In: *Families in Society* 98.2, pp. 121–125. DOI: [10.1606/1044-3894.2017.98.17](https://doi.org/10.1606/1044-3894.2017.98.17).

Appendix

1 EUROCROPS: The Largest Harmonized Open Crop Dataset Across the European Union

Maja Schneider, Tobias Schelte, Felix Schmitz and Marco Körner (2023). 'EuroCrops: The Largest Harmonized Open Crop Dataset Across the European Union'. In: *Scientific Data* 10.1, p. 612. DOI: [10.1038/s41597-023-02517-0](https://doi.org/10.1038/s41597-023-02517-0)

scientific data



OPEN

DATA DESCRIPTOR

EuroCrops: The Largest Harmonized Open Crop Dataset Across the European Union

Maja Schneider , Tobias Schelte , Felix Schmitz  & Marco Körner


EuroCrops contains geo-referenced polygons of agricultural croplands from 16 countries of the *European Union (EU)* as well as information on the respective crop species grown there. These semantic annotations are derived from self-declarations by farmers receiving subsidies under the *common agricultural policy (CAP)* of the *European Commission (EC)*. Over the last 1.5 years, the individual national crop datasets have been manually collected, the crop classes have been translated into the English language and transferred into the newly developed *hierarchical crop and agriculture taxonomy (HCAT)*. EuroCrops is publicly available under continuous improvement through an active user community.

Background & Summary

As the world's population continues to grow and global climate change becomes increasingly apparent, enhancing the efficiency and resilience of agriculture at both the local and global level is a crucial challenge for humanity's future. Recent developments in satellite-based *Earth observation (EO)* have provided us with the ability to observe and analyse the processes occurring on the Earth's surface in near real-time. By leveraging machine learning and artificial intelligence, we can extract valuable insights from these enormous volumes of high-quality and information-rich data, which can inform the development of functional process models for the monitoring of agricultural crops and the design of future applications. For example, the activity of these vegetation stands could be monitored and deviations from the expected progression, and thus the expected crop yields, could be detected. Based on this information, farmers would be able to initiate countermeasures at an early stage. This would make a decisive contribution to food security, representing one of the central *sustainability development goals (SDGs)* stated by the *United Nations (UN)*. However, these possibilities are massively limited by the insufficient availability of qualitative reference data, which are necessary for the creation of functional process models on the basis of such Earth observation data.

The EuroCrops project aims to show how this gap can be filled by compiling administrative data assessed in the context of agricultural subsidy control in the *European Union (EU)* area. Therefore, publicly available *Land Parcel Identification System (LPIS)* data, the essential part of the spatial information used to support *Integrated Administration and Control System (IACS)* applications under the *common agricultural policy (CAP)*, is individually collected from the member states. It contains georeferenced blocks of agricultural parcels that have been identified and are eligible for EU aid application. This data usually shows the main crop for a certain year as the subsidy is granted with respect to that.

A first pilot project¹ exemplified the process compiling a dataset from that type of data. For this purpose, we collected geo-referenced crop datasets from three countries within Europe, harmonised the data by translating the crop names and developed an hierarchical structure to order the occurring crops. Finally, the crop labels were paired with the corresponding Sentinel-2 EO data and we released the TinyEuroCrops² dataset publicly via the repository of the Technical University of Munich. Despite faced with some challenges, we soon realised that the dataset gained its popularity not due to the satellite data, but due to the fact that we also published the geo-referenced field polygon vector data together with the harmonised information of which crop species were cultivated there for a certain year. Having this data prepared in one reconciled format, language, and centrally available across borders and not just on a national level sparked the discussions about its broad applicability in various domains. The fact that this data has been prepared in a joint standardised format and language and that it is centrally available across borders and not only at national level has triggered discussion about its broad applicability in various areas. The research questions related to the analysis of agricultural diversity and food

Technical University of Munich (TUM), TUM School of Engineering and Design, Munich, 80333, Germany. e-mail: maja.schneider@tum.de

security in Europe were one of the reasons for the popularity of the dataset, leading to the motivation to extend them spatially and later also temporally.

In this article, we present and describe the first spatially extended EuroCrops vector dataset. For this release, we manually collected the raw crop declaration data from 16 EU countries, which was made available and distributed across multiple platforms and servers. In light of previous studies, e.g., BreizhCrops³, ZueriCrop⁴, and CropHarvest⁵, the key objectives of EuroCrops lie in the extension of both the variability of crop species classes to be represented and the geographical scale of the considered regions. After translating the textual declarations data, we developed a new version of our *hierarchical crop and agriculture taxonomy (HCAT)*¹ in order to organize all crops that are cultivated within the EU into a common hierarchical representation scheme. The process of this development is visualised in Fig. 1 and will be further explained in the methods section.

By being able to analyse agricultural data at this expanded spatial scale, which extends from Sweden to Portugal, we hope to enable researchers to carry out their work across borders and gain new insights. EuroCrops is ongoing and will be extended on a regular basis. We are putting effort into increasing the spatial and temporal coverage of the dataset as well as the preparation of analysis-ready data by combining it with Sentinel-2 data. Updates will be provided on Zenodo and the GitHub repository associated with the project.

Methods

In order to compile the presented dataset, several iterative steps had to be performed, which can roughly be grouped into *data collection*, *harmonisation* and *validation*, denoted as **A**, **B** and **C** in Fig. 1 and will be further described in the next subsections. Data obtained from each member state of the EU has to undergo the entire procedure, sometimes even multiple times, as indicated by the stacked layers in Fig. 1 and arrows going from each country's Update HCAT process back to the beginning and the Automatic mapping to HCAT for the individual dataset. This recurring loop is the main reason for the exponentially increasing amount of manual work that was necessary for the creation of the dataset and required careful deliberation on the right moment for cutting the development of HCAT.

A. Data collection. As EuroCrops consists of multiple smaller datasets, the data collection itself plays an integral role. This paper will focus on the practical part of that process, whereas an in-depth analysis of the challenges of creating a transnational dataset is described in more detail by Schneider *et al.*⁶.

Generally, we identified four ways of data acquisition: Firstly, many countries publish national crop data on the webpage of the respective ministry or agency responsible for agricultural, food or rural topics. Some countries instead offer a national geoportal, distributing different kinds of geodata specifically or, as another mean of distribution, publishing geodata on an international level, e.g. via INSPIRE⁷ or data.europa.eu⁸. Lastly, if the data is not openly distributed on a webpage or geoportal, we reached out personally to ministries or agencies and asked for the data directly. Most of the national datasets used in the EuroCrops project were collected from national ministry webpages or geoportals as listed in Tables 1, 2 respectively, mostly made available as *ESRI shapefiles*, *GeoJSON*, or *GeoPackage (GPKG)*. Nonetheless, some data can only be accessed via a *web feature service (WFS)* implemented in a *geographic information system (GIS)*, allowing the user to display the desired data and save it in a chosen file format. The other means of data access are shown in Tables 3, 4. Figure 3 puts all this information into context, gives an overview of the available datasets, and indicates from where the data originates. Countries marked yellow in Fig. 3 indicate only partial availability of crop data for the respective country. In order to give a better understanding of the original raw datasets we got from the countries, we visualised a small fraction of the data from North-Rhine Westphalia (Germany) in Fig. 2 with coloured geo-referenced agricultural parcel polygons. Table 5 gives an impression of how the corresponding original raw attribute table looks like. Each row entry describes the crop species that has been cultivated on the associated parcel.

From all red coloured countries in Fig. 3 it was not possible to obtain publicly available data. There are several reasons for that, such as *General Data Protection Regulation (GDPR)* issues, a missing incentive to publish the datasets and sometimes no response over years from the responsible authorities. However, as these types of data collections have recently been declared high-value datasets by the *European Commission (EC)* (<https://digital-strategy.ec.europa.eu/en/news/commission-defines-high-value-datasets-be-made-available-re-use>), we expect a change towards a more open publishing culture in the future.

In the following paragraphs, all individual sources for the available datasets are presented. For each contributing country the data source, available years, coverage, licence, and format are described and referenced. By doing so, we aspire to give the research community a tool to discover and access raw data faster and more reliably.

Austria. The dataset for Austria comprises a vast range of years, spanning from 2015 to 2021. Moreover, the whole territory of the country is covered without any regional omissions. Crop classes are defined very detailed with an approximate number of 200 classes. The files were made available in GPKG format via two platforms, the European “data.europa.eu”⁹ and “data.gv.at”¹⁰, a platform that distributes data of the public sector in Austria for further analysis and development. However, both platforms receive the datasets from “Agrarmarkt Austria”, which is a public geodata office. As such, data is published free of charge under the Creative Commons Licence CC-BY-AT 4.0. In the course of the EuroCrops project, the dataset of 2021 was harmonised for Austria.

Belgium. Due to the federal structure of Belgium, the data is split into two sets covering the regions of Flanders and Wallonia. Not only is the data published via different platforms, its structure also differs heavily between the two regions.

The data for Flanders¹¹ is published by the Department of Agriculture and Fishery on its website as shapefiles. It is anonymous and can be used freely. Additionally, a word document explaining the current state of the

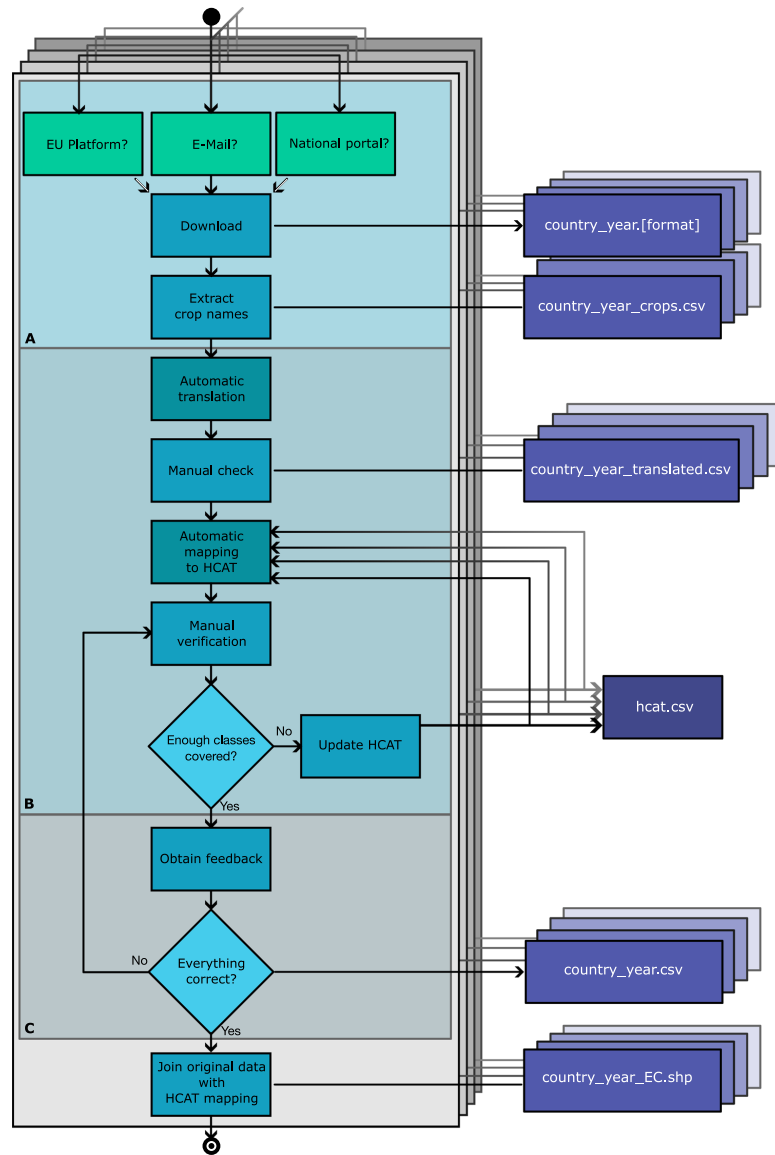


Fig. 1 The process of constructing the EuroCrops dataset. Each layer represents the process for one country with the three stages of development: **A. Data Collection**, **B. Harmonisation**, **C. Validation and Feedback**. Each stage has one or more outputs, indicated in purple, and with `country` and `year` being replaced accordingly. Only the `hcat.csv` exists once across all country-specific processes and gets gradually updated in each harmonisation step. While automatic sections exist, a manual check is required each time, making the progress in total heavily dependent on work that has to be done by hand.

data as well as the abbreviations that occur in the attribute table of the shapefiles is available in Flemish language. The crop classes are differentiated very precisely with an approximate number of 275 classes. Datasets are available for the years 2019, 2020 and 2021.

The datasets for Wallonia are published by the Geoportail of Wallonia (<https://geoportail.wallonie.be/catalogue-donnees-et-services>) as shapefiles, but a registration is required. With an approximate number of

| Country (state) | National agency | URL | Format |
|--------------------|--|---|-----------------|
| Belgium (Flanders) | Department of Agriculture and Fisheries (Departement Landbouw & Visserij) | see Departement Landbouw en Visserij ¹¹ | shapefile, GPKG |
| Croatia | Agency for Payments in Agriculture, Fisheries and Rural Development | https://www.aprrr.hr/prostorni-podaci-servisi/ | GPKG |
| Denmark | Danish Agricultural Agency | https://landbrugsgeodata.fvm.dk/ | shapefile |
| Finland | Finish Food Authority | https://kartta.paikkatietoikkuna.fi/ | WFS |
| Latvia | Rural Support Service Republic of Latvia (Lauku atbalsta dienests) | https://www.lad.gov.lv/lv/lauku-registra-dati | WFS |
| Netherlands | Ministry of economic affairs and climate (Ministerie van Economische Zaken en Klimaat) via: PDOK platform (Publieke dienstverlening op de kaart) | https://www.pdok.nl/introductie/-/article/basisregistratie-gewaspercelen-brp | WFS |
| Portugal | Portuguese Finance Institute of Agriculture and Fisheries (Instituto de Financiamento da Agricultura e Pescas) | https://www.ifap.pt/isip/ows/ | WFS |
| Slovenia | Ministry of Agriculture, Forestry and Food (Ministrstvo za Kmetijstvo, Gozdarstvo in Prehrano) | https://rkg.gov.si/vstop/ | shapefile |

Table 1. National (ministry) website: The majority of the EuroCrops data sources were websites, usually hosted by the respective ministry or agency. These websites are usually in the national language, without any English translation which makes the discoverability and accessibility of the data laborious for international researchers.

150 classes the crop classification of Wallonia is still quite precise, even though the Flemish data is more detailed. On the other hand, a wider time period is captured by the Wallonian datasets, covering all years since 2015.

So far only the Flemish data for the year 2021 got harmonised in the course of the EuroCrops-Project.

Croatia. The Croatian data is distributed in GPKG format via a platform managed by the Agency for Payments in Agriculture, Fisheries and Rural Development (<https://www.aprrr.hr/prostorni-podaci-servisi/>), where an abundant sequence of years is available ranging from 2011 to 2021. Due to translation difficulties, we obtained the data directly from the Paying Agency with the rights to include it in EuroCrops. While all regions of the country are covered by the dataset, its differentiation between 14 crop classes turns out to be rather coarse. For EuroCrops, the data of 2020 was harmonised.

Denmark. The dataset of Denmark comprises of only the mainland. The Faroe Islands and Greenland are not included. However, with an approximate number of 300 classes, the Danish crop taxonomy is very detailed. Datasets are available since the year 2017. The data is available as shapefiles provided by the Danish Agricultural Agency (<https://landbrugsgeodata.fvm.dk/>). All the data provided is considered open data, which means it can be openly used and distributed. The Danish data of 2019 was harmonised throughout the course of the EuroCrops Project.

Estonia. The Estonian dataset¹² is made available under the “Autorile viitamine-Jagamine samadel tingimustel 3.0 Eesti” which corresponds to a CC-BY-SA licence. Thus, there are no limitations to public access. It can be acquired via the INSPIRE Geoportaal as WFS. When accessing the data via a WFS URL in a GIS, the dataset can be transformed and saved as GeoJSON for example. It covers all of Estonia but only for the current year. Thus, data from 2021 was harmonised. However, the crop differentiation is very precise, leading to a high number of ca. 150 classes.

Finland. The Finnish dataset covers all provinces of the country. Data is available for the years 2020 and 2021. However, none of the years got harmonised yet, as Finland provided its datasets very late after the harmonisation process was already completed. The data differentiates between 200 classes roughly, which enables a very precise crop classification. The Finish Food Authority distributes the data via a WFS (<https://kartta.paikkatietoikkuna.fi/>) under the Creative Commons Licence BY 4.0. Consequently, the datasets were implemented into a GIS and saved as shapefile.

France. France publishes national geodata as open licence on the “data.gouv.fr” platform¹³ as GPKG- and shapefiles. While the central point of distribution makes it easy to discover and access the data, the fact that each region has its own sub-dataset makes the platform barely usable for someone who needs the entirety of the French data. Luckily, there is a second (unofficial) server (data.cquest.org/registre_parcellaire_graphique/2018/) that hosts a combination of all these national datasets in shapefile format. Additionally, an excel sheet is available, containing the descriptions of all crop abbreviations used in the datasets. The class differentiation is moderate. Approximately 70 crop classes are distinguished. In the course of the project, datasets were downloaded for the years spanning from 2016 to 2019, of which the file for 2018 was harmonised. The data covers not only the French mainland but also overseas territories.

Germany. Due to the federal structure of Germany datasets are not published on a national level, but by each federal state (“Bundesland”) individually. Two datasets were acquired: One covers Lower Saxony (<https://sla>).

| Country (state) | Portal | URL | Format |
|----------------------------------|---|---|-----------|
| Austria | data.gv.at | see Agrarmarkt Austria ¹⁰ | GPKG |
| Belgium (Wallonia) | Géoportail de la Wallonie | https://geoportail.wallonie.be/catalogue-donnees-et-services | shapefile |
| France | data.gouv.fr | see Agence de services et de paiement (ASP) ¹³ | shapefile |
| Germany (North Rhine-Westphalia) | OpenGeodata.NRW | see Landwirtschaftskammer NRW ¹⁴ | shapefile |
| Germany (Lower Saxony) | Landentwicklung und Agrarförderung Niedersachsen-Portal | https://sla.niedersachsen.de/landentwicklung/LEA/ | shapefile |
| Lithuania | geoportal.lt | see Nacionalinė mokėjimo agentūra prie Žemės ūkio ministerijos ¹⁶ | shapefile |
| Spain | "Sicpac" portal for each Autonomous Community, e.g. Navarra | https://flicscartografia.navarra.es/2_CARTOGRAFIA_TEMATICA/2_6_SIGPAC/ | |

Table 2. National geoportal: Some countries or regions actively participate in Europe's open data initiative and publish their crop data on a national geoportal. The goal of these portals is to make data available to the public sector and lower the entry barrier to letting citizens actively participate.

| Country | Authority | Format |
|----------|---------------------------------------|-----------|
| Slovakia | National Agricultural and Food Centre | shapefile |
| Sweden | The Swedish Board of Agriculture | shapefile |

Table 3. Direct contact: Data from Slovakia and Sweden was directly sent to the project members by a contact person in the respective country.

| Country | Portal | Format |
|-----------|---------------------------------|-----------|
| [Austria] | data.europa.eu ⁹ | GPKG |
| Estonia | INSPIRE Geoportal ¹² | WFS |
| Romania | INSPIRE Geoportal ¹⁷ | shapefile |

Table 4. International Platform This table lists all the countries which data was acquired through an international platform. The data for Austria would be available via "data.europa.eu", but for EuroCrops we downloaded it directly from the national website.

[niedersachsen.de/landentwicklung/LEA/](https://sla.niedersachsen.de/landentwicklung/LEA/)) and another one North Rhine-Westphalia¹⁴. Both datasets depict the crop situation of 2021 and have a very high class precision, distinguishing between ca. 240 crop classes. Both files are distributed as shapefiles, one on the online platform for Rural Development and Agricultural Promotion of Lower Saxony, the other one on the geoportal of North Rhine-Westphalia. Both datasets are published under "data licence Germany - attribution - Version 2.0". For the sake of completeness, it is worth noting that Brandenburg also published its data¹⁵, but has not been included into EuroCrops yet.

Latvia. The Rural Support Service of Latvia (<https://www.lad.gov.lv/lv/lauku-registra-dati>) provides a WFS, which can be used to implement and convert the Latvian files to `GeoJSON` or shapefile in a GIS. The data is open so there are no publishing restrictions. The files cover the whole territory of the country and are available for 2021 and 2022. The file for 2021 got harmonised in the course of the EuroCrops Project. The class precision is very high, differentiating between 150 crop types approximately.

Lithuania. The crop parcels of Lithuania¹⁶ are available as shapefiles for the year 2021 covering the whole territory of the country. Consequently, data got harmonised for the aforementioned year. The file differentiates between 24 crop classes only. However, the chosen classes are precise. Datasets of a similar low number of classes normally assign very general crop terms to the classes (i.e. vineyard, citrus fruits, grassland). In the case of Lithuania, the crop types assigned to the classes are very specific. Thus, the class precision can be defined as medium, despite its low number of actual classes. The data is published via Geoportal.lt under their own copyright, but it requires registration.

Netherlands. The Dutch Ministry of Economic Affairs and Climate distributes datasets via a WFS on the platform PDOK (<https://www.pdok.nl/introductie/-/article/basisregistratie-gewaspercelen-brp->). The files comprise only the mainland of the Netherlands; overseas territories are not included. The class precision is very high, encompassing around 320 different plant categories. So far data is only available for the years 2020 and 2021, of which the file for 2020 was harmonised for EuroCrops. The datasets fall under the CC0 1.0 licence category which does not impose any limitations to public access.

www.nature.com/scientificdata/

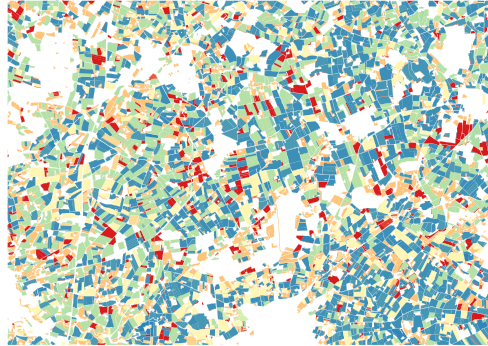


Fig. 2 Exemplified raw input data with the corresponding attribute table shown in Table 5. This selection shows parts of the North Rhine-Westphalian (Germany) dataset¹⁴ with each crop class being coloured differently. The data consists of geo-referenced polygons which indicate the field borders and hold information about the grown crop for a certain year.

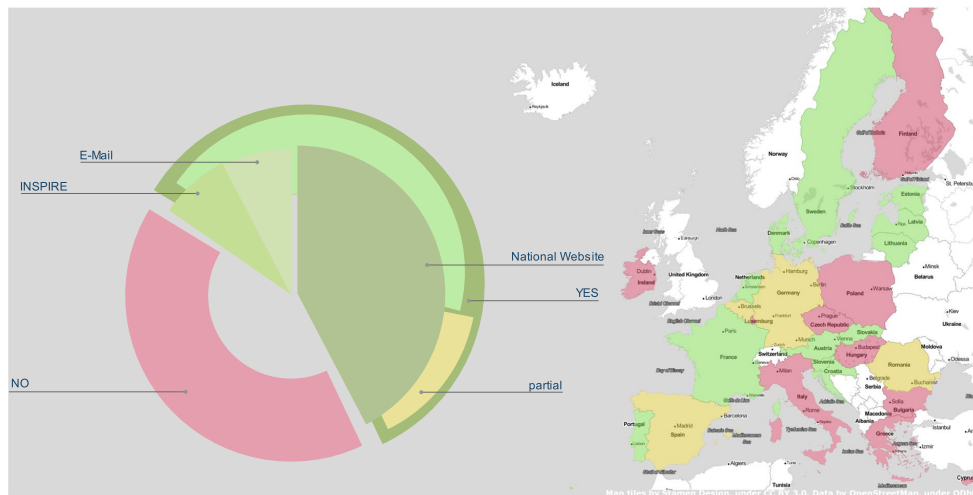


Fig. 3 This diagram shows the data availability, coverage and access across the EU. The map on the right indicates whether data from a certain country is part of EuroCrops. Green indicates a full country participation, yellow are the member states with only partial coverage and data from red countries was not available at the time of the development of the dataset. Finland for instance released the data later and is still covered in the section “Data Records”, but is not part of EuroCrops. On the left, the pie chart breaks down several analyses: Generally, red and green shades tell what fraction of data in the EU is available, while indicating with bright green and yellow full or partial coverage, as already on the map. The green pie segments on top visualise where the data was originally obtained from. This shows, similar to Table 1 that the majority of the data originates from national websites, including geoportals.

Portugal. The Portuguese datasets are available since 2017, with the file for 2021 harmonised. Data since 2020 covers the complete national territory of Portugal. Contrarily, the files for 2017, 2018 and 2019 are split up into regional territories, which had to be merged in a first step. Moreover, some of the Portuguese regions are missing whereas the national datasets provide a complete and uniform depiction of Portuguese crop cover. Furthermore, the class precision differs between the regional and the national datasets. The crop differentiation is moderate for the regional sets with ca. 50 to 150 classes, whereas it is more precise for the national datasets with more than 200 classes. The files can be accessed via a WFS (<https://www.ifap.pt/isip/ows/>) provided by the Portuguese Finance Institute of Agriculture and Fisheries and is usable without legal restrictions.

| ID | FLIK | AREA_HA | CODE | CODE_TXT | USE_CODE | USE_TXT | WJ | DAT_BEARB |
|---------|------------------|---------|------|----------------------------|----------|---------------|------|------------|
| 4597509 | DENWLI0543050566 | 2.1808 | 411 | Silomais (als Hauptfutter) | AF | Ackerfutter | 2021 | 2021/03/01 |
| 4597510 | DENWLI0543051616 | 1.5319 | 459 | Grünland (Dauergrünland) | GL | Dauergrünland | 2021 | 2021/03/01 |
| 4597641 | DENWLI0542022516 | 1.0293 | 480 | Streuobst mit DGL-Nutzung | GL | Dauergrünland | 2021 | 2021/03/02 |
| 4597657 | DENWLI0541093620 | 2.4966 | 459 | Grünland (Dauergrünland) | GL | Dauergrünland | 2021 | 2021/03/02 |
| 4597810 | DENWLI0540163053 | 1.162 | 121 | Winterroggen | GT | Getreide | 2021 | 2021/03/04 |

Table 5. After downloading the respective national raw datasets, we first examined the attribute tables and extracted the values of the columns representing the cultivated crops. In the given example, showing raw data from North Rhine-Westphalia (Germany)¹⁴, multiple columns representing the cultivated crops can be determined. Thus a selection had to be made. In this case, values from the “CODE_TXT” column were translated and matched with the occurring classes in HCAT. Each time we discovered a class that was not represented in the taxonomy yet, we included it and started the harmonisation process again. The corresponding vector data to this file is illustrated in Fig. 2 and the same attribute table enriched with the EuroCrops columns is shown in Table 6.

Romania. Romania officially does not yet publish crop data but is, according to the *Agenția de Plăți și Intervenție pentru Agricultură*, actively working towards it. We therefore decided to add an unlicensed, coarse and only regional land cover dataset¹⁷ into EuroCrops in order to give an incentive and an idea of how Romanian data would be integrated in the future.

Slovenia. The Slovenian dataset covers the territory of the whole country and the years 2019, 2020 and 2021. The file for 2021 got harmonised. The class precision is high, with approximately 150 different crop classes. The files are distributed as shapefiles at the website of the Ministry of Agriculture, Forestry and Food (<https://rkg.gov.si/vstop/>). Additionally, two text files are published which describe the crop codes assigned to the plants with one file being in Slovenian language, the other one in English. All data is made publicly available without use restrictions, however, citing the source is required.

Slovakia. Slovakian data is available for the years 2020, 2021 and 2022. The datasets cover all regions of the country. The file depicting the crop situation in 2021 was harmonised. The class precision is very high, differentiating between roughly 170 crop types. The data was sent directly to the project members via e-mail by the Slovakian Agricultural Paying Agency with the permission to include it into EuroCrops.

Spain. Spain distributes data under the licence CC BY 4.0 separately for each of its autonomous communities where each one has their own website. The crop parcel data can be downloaded there as a shapefile in most cases. The Navarra dataset (https://filescartografia.navarra.es/2_CARTOGRAFIA_TEMATICA/2_6_SIGPAC/) for 2021 got harmonised. However, the data is very coarse, differentiating between 21 classes only.

Sweden. GeoJSON files covering the crop parcels of all of Sweden for the years 2020 and 2021 were sent by a contact person at the Swedish Board of Agriculture to the project members by email. The files have a medium class precision distinguishing between ca. 80 classes, are published under the CC BY 4.0 licence and data depicting the crop situation in 2021 got harmonised.

B. Harmonising country-specific crop classes. After collecting the data and extracting the set of country-specific crop classes from the attribute tables, we initiated the harmonisation process. This step is necessary because the crop names from each country usually come in the national language of the member states and without standardised codes, as shown in Table 2.

Instead of working with the entire attribute table, we worked with a table showing the name and code of a certain crop class per row together with its absolute and relative occurrence in the dataset. In Fig. 1, that file, preserving original crop class name and code, is denoted as `country_year_crops.csv`.

Following this, the automatic translation starts the process of harmonising the given original crop class names into the HCAT taxonomy. Therefore, multiple steps had to be performed: The file `country_year_translated.csv` arises from the translation of the crop classes into English. Despite the access to modern translation programmes, we were not able to automate this part end-to-end, as country-specific agricultural terms seem to cause mistranslations across all common translators. By correcting the translations manually, we hope to bridge that gap and make the dataset as reliable as possible. Similarly, mapping the translations to HCAT was only possible to perform automatically to a certain degree and required manual checks (see Fig. 1) as well. This was again caused by the diversity of the crop classes declared by the member states.

Within this process of manual translation and matching, we were able to catch most of the missing classes in our growing taxonomy. The iterative updates of HCAT helped in the detailed classification of delivered datafiles by the countries, but also shed light on relevant and focus areas within the taxonomy. To the matching HCAT name, we also added the corresponding HCAT code, which embeds the hierarchy of the taxonomy. This way, we enriched the country-specific original crop name and code with our HCAT name and code and the absolute and relative occurrence in a country.

| ID | ... | CODE_TXT | USE_CODE | USE_TXT | WJ | DAT_BEARB | EC_trans_n | EC_hcat_n | EC_hcat_c |
|---------|-----|----------------------------|----------|---------------|------|------------|---------------------------------------|--------------------------------|------------|
| 4597509 | ... | Silomais (als Hauptfutter) | AF | Ackerfutter | 2021 | 2021/03/01 | Silage maize (as staple feed) | green_silo_maize | 3301090400 |
| 4597510 | ... | Grünland (Dauergrünland) | GL | Dauergrünland | 2021 | 2021/03/01 | Grassland (permanent grassland) | pasture_meadow_grassland_grass | 3302000000 |
| 4597641 | ... | Streuobst mit DGL-Nutzung | GL | Dauergrünland | 2021 | 2021/03/02 | Orchards with Permanent grassland use | orchards_fruits | 3303010000 |
| 4597657 | ... | Grünland (Dauergrünland) | GL | Dauergrünland | 2021 | 2021/03/02 | Grassland (permanent grassland) | pasture_meadow_grassland_grass | 3302000000 |
| 4597810 | ... | Winterroggen | GT | Getreide | 2021 | 2021/03/04 | Winter rye | winter_rye | 3301010301 |

Table 6. This table shows an example of a final EuroCrops data attribute table. While the original columns as shown in Table 5 remains the same (“FLIK”, “AREA_HA” and “CODE” have been abbreviated in this print), three additional attributes were added: Firstly, “EC_trans_n” is the direct translation of the crop name in its original language. Then, correspondingly, “EC_hcat_n” is the matched name of that particular crop in *HCAT*. These names are all lowercase and with underscores to make them easier to process automatically. Lastly, the column “EC_hcat_c” shows the HCAT code that puts the HCAT names into a hierarchical structure. A more detailed explanation of HCAT is presented in the publication by Schneider *et al.*⁴.

Hence, we are able to visualise the number of instances of certain crop classes and compare the occurrences with those from other countries for general diversity analysis and taxonomy class updates. The preliminary file is stored in a `country_year.csv` after positive assessment during working step C.

C. Community work: content validation and feedback incorporation. The largest expertise on country-specific crop classes still lies with the respective countries, driving to the decision to keep them onboard during the validation phase of the project. Therefore, we asked all countries during the end phase of our pipeline if our translations and mappings seemed reasonable. Out of 16 countries, we received feedback from seven who double-checked and reviewed our work. While this increased the quality of the dataset, it also started another loop in the harmonisation block, which is visualised in Fig. 1 as the arrow going from *Everything correct?* to *Manual verification*. Eventually, we uploaded the first version of the dataset on our university-owned data-sharing platform and set up a GitHub repository (<https://github.com/maja601/EuroCrops>) for the community to have a first look. This resulted into several opened issues and pull requests where improvements to the mappings were suggested. Each time we were content with a version of the mapping, we manually joined the original dataset with our mapping and saved it as a shapefile. This led to one shapefile for each country and five successive versions of the dataset incorporating the proposed changes from GitHub. One exemplary attribute table of such a shapefile is shown in Table 5. All of the versions were individually uploaded to Zenodo¹⁸, which now officially tracks the versions with a *Digital Object Identifier (DOI)*.

Data Records

The EuroCrops dataset is currently published as individual country- or region-covering shapefiles and hosted on Zenodo¹⁸ with dynamic updates available on GitHub (<https://github.com/maja601/EuroCrops>). Therefore, all individual file types that are required (`.shp`, `.shx`, `.dbf`) and optional (`.cpg`, `.prj`) to form a shapefile are zip compressed as one sub-dataset directory and can be downloaded individually. This way, researchers that are only interested in a specific area can make use of a selection of EuroCrops without having to download everything. The naming convention for the individual files is the country name in ISO-3166 Alpha-2 format with an optional regional identifier and the year for which the data has been harmonised. The attribute tables of the original shapefiles which have been introduced in Section A. **Data Collection** are not altered throughout the process, but amended by the three columns `EC_trans_n`, `EC_hcat_n`, `EC_hcat_c`, representing the translated crop name, the HCAT name and the HCAT code respectively as shown in Table 6.

Technical Validation

Regarding the correctness of the underlying original data, it is important to stress that self-declarations build the basis of the input. From official site, the *in-situ* controls act as a validation instance to these declarations, but these are just sparse samples and would never be able to cover the entire area. One approach to actually validate the original data on a bigger scale was introduced by Gounari *et al.*¹⁹, but this would exceed the project tasks. On our side, we concentrated on a valid harmonisation of the entire dataset. The validation of the content itself was already discussed in the methods section: We incorporated the knowledge from the respective authorities and updated the mappings based on feedback from the community.

| original_code | original_name | translated_name | HCAT2_name | HCAT2_code |
|---------------|----------------------------|---------------------------------|--------------------------------|------------|
| 459 | Grünland (Dauergrünland) | Grassland (permanent grassland) | pasture_meadow_grassland_grass | 330200000 |
| 411 | Silomais (als Hauptfutter) | Silage maize (as staple feed) | green_silo_maize | 330109040 |

Table 7. In our aforementioned GitHub repository, we publish for each country a so-called mapping file. This file contains the set of occurring crops in one original file, together with its translation and the corresponding HCAT name and code. Note, that even though it says HCAT2 in the column names, it is the same as the previously mentioned HCAT. As the initial, prototyped taxonomy is not used any more. The shown example is an extraction of the sub-dataset which is already presented in Fig. 2 and Tables 5, 6. The entire file is available on GitHub and could for example be used to translate and map a dataset from North Rhine-Westphalia (Germany)¹⁴ of another year than 2021.

Usage Notes

The data is currently published as one shapefile per country on Zenodo¹⁸ which can, for instance, be opened with QGIS²⁰. On Zenodo, EuroCrops version 6 is associated with this peer-reviewed article. The corresponding, dynamically updated mapping files on GitHub (https://github.com/maja601/EuroCrops/tree/main/csvs/country_mappings) are in CSV format with the structure found in Table 7. In order to use data from a year that has not been harmonised within EuroCrops, it is possible to join the mapping file of a country with the raw vector data file which can be found on the provided national platforms. By using the correct column in the original dataset, which is indicated in the wiki (<https://github.com/maja601/EuroCrops/wiki>) entry under “Attribute Table” for each country, also other datasets can be harmonised. This might lead to some missing crop types, as our taxonomy only holds the crop classes occluding in the stated sub-datasets, but we assume that the majority of the crops should be covered.

Code availability

For this study, no custom code was generated.

Received: 17 January 2023; Accepted: 30 August 2023;

Published online: 11 September 2023

References

- Schneider, M., Broszeit, A. & Körner, M. EuroCrops: A pan-european dataset for time series crop type classification. In Soille, P., Loekken, S. & Albani, S. (eds.) *Proceedings of the Conference on Big Data from Space (BiDS)*, <https://doi.org/10.2760/125905> (Publications Office of the European Union, 2021).
- Schneider, M. & Körner, M. TinyEuroCrops, <https://doi.org/10.14459/2021MP1615987> (2021).
- Rußwurm, M., Pelletier, C., Zollner, M., Lefèvre, S. & Körner, M. Breizhcrops: A time series dataset for crop type mapping. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences XLIII-B2-2020*, 1545–1551, <https://doi.org/10.5194/isprs-archives-XLIII-B2-2020-1545-2020> (2020).
- Turkdoglu, M. O. et al. Crop mapping from image time series: Deep learning with multi-scale label hierarchies. *Remote Sensing of Environment* **264**, 112603, <https://doi.org/10.1016/j.rse.2021.112603> (2021).
- Tseng, G., Zvonkov, I., Nakalembe, C. L. & Kerner, H. CropHarvest: A global dataset for crop-type classification. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)* (2021).
- Schneider, M., Marchington, C. & Körner, M. Challenges and Opportunities of Large Transnational Datasets: A Case Study on European Administrative Crop Data. *Workshop on Broadening Research Collaborations in ML (NeurIPS 2022)* <https://doi.org/10.48550/arXiv.2210.07178> (2022).
- INSPIRE Geoportal Team. INSPIRE Geoportal. <https://inspire-geoportal.ec.europa.eu/> (2022).
- European Commission. data.europa.eu - The official portal for European data. data.europa.eu <https://data.europa.eu> (2022).
- Agrarmarkt Austria. InVeKoS References Austria 2021. data.europa.eu http://data.europa.eu/88u/dataset/ama_invekosreferenzensterreich2021 (2021).
- Agrarmarkt Austria. INVEKOS Schläge Österreich 2021. [data.gv.at](https://www.data.gv.at) <https://www.data.gv.at/katalog/dataset/fa18db4f-a880-452b-bcbf-e4c0a88cb5d5> (2021).
- Departement Landbouw en Visserij – Landbouwcijfers. Landbouwgebruikspercelen. <https://landbouwcijfers.vlaanderen.be/opengeodata-landbouwgebruikspercelen> (2021).
- Põllumajanduse Registrite ja Informatsiooni Amet. Geospatial Aid Application Estonia Agricultural parcels. *INSPIRE Geoportal* <https://inspire-geoportal.ec.europa.eu/results.html?country=ee&view=details&theme=none> (2021).
- Agence de services et de paiement (ASP). Registre parcellaire graphique (RPG): contours des parcelles et ilots culturaux et leur groupe de cultures majoritaire. [data.gouv.fr](https://www.data.gouv.fr/en/datasets/registre-parcellaire-graphique-rpg-contours-des-parcelles-et-ilots-culturaux-et-leur-groupe-de-cultures-majoritaire/) <https://www.data.gouv.fr/en/datasets/registre-parcellaire-graphique-rpg-contours-des-parcelles-et-ilots-culturaux-et-leur-groupe-de-cultures-majoritaire/> (2018).
- Landwirtschaftskammer NRW. LWK-TSCHLAG. [opengeodata.nrw.de](https://www.opengeodata.nrw.de/produkte/umwelt_klima/bodennutzung/landwirtschaft/) https://www.opengeodata.nrw.de/produkte/umwelt_klima/bodennutzung/landwirtschaft/ (2021).
- Ministerium für Landwirtschaft, Umwelt und Klimaschutz des Landes Brandenburg (MLUK). Agrarantragsdaten. *GEOBROKER* <https://geobroker.geobasis-bb.de/gbss.php?MODE=GetProductInformation&PRODUCTID=996f8fd1-c662-4975-b680-3b611fcb5d1f> (2021).
- Nacionalinė mokėjimo agentūra prie Žemės ūkio ministerijos. Lietuvos Respublikos teritorijos Žemės ūkio naudmenų ir pasėlių plotų, auginamųjų kultūrų duomenų rinkinys. *geoportal.lt* <https://www.geoportal.lt/geoportal/nacionaline-mokejimo-agentura-prie-zemes-ukio-ministerijos#savedSearchId=772172A4-6719-48BD-8DDC-5DEEFB27DE74&collapsed=true> (2021).
- Ministerul Dezvoltării Regionale și Administrației Publice. Harta tematica a acoperirii terenurilor în aria transfrontalieră România-Bulgaria. [data.europa.eu](https://data.europa.eu/data/datasets/092425a1-90c6-4461-b1a6-6f5b0f72748f?locale=ro) <https://data.europa.eu/data/datasets/092425a1-90c6-4461-b1a6-6f5b0f72748f?locale=ro> (2017).
- Schneider, M. & Körner, M. EuroCrops. *Zenodo* <https://doi.org/10.5281/zenodo.7476474> (2022).
- Gounari, O., Karakizi, C. & Karantzalos, K. Filtering LPIS data for building trustworthy training datasets for crop type mapping: A case study in Greece. *The International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences XLIII-B3-2022*, 871–877 (2022).
- QGIS Development Team. QGIS Geographic Information System. www.qgis.org.

Acknowledgements

The authors and the EuroCrops project receive funding from the German *Federal Ministry for Economic Affairs and Climate Action* on the basis of a resolution of the German Bundestag under reference 50EE1908 and from the European Union's *Horizon 2020* research and innovation programme under grant agreement No 101004112.

Author contributions

M.S. leads the EuroCrops project, identified data sources, created HCAT, obtained feedback from authorities and compiled the published shapefiles. T.S. obtained and documented the individual datasets from the data sources. F.S. translated the crop classes and analysed the individual datasets. M.S., T.S. and F.S. verified the crop translations and mappings to HCAT. M.K. was involved in the design of the concept and supervised the project. All authors reviewed the manuscript.

Funding

Open Access funding enabled and organized by Projekt DEAL.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to M.S.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2023

2 Harnessing Administrative Data Inventories to Create a Reliable Transnational Reference Database for Crop Type Monitoring

© 2022 IEEE. Reprinted, with permission, from Maja Schneider and Marco Körner (2022). 'Harnessing Administrative Data Inventories to Create a Reliable Transnational Reference Database for Crop Type Monitoring'. In: *IGARSS 2022 - 2022 IEEE International Geoscience and Remote Sensing Symposium*, pp. 5385–5388. DOI: [10.1109/IGARSS46834.2022.9883089](https://doi.org/10.1109/IGARSS46834.2022.9883089)

HARNESSING ADMINISTRATIVE DATA INVENTORIES TO CREATE A RELIABLE TRANSNATIONAL REFERENCE DATABASE FOR CROP TYPE MONITORING

Maja Schneider* and Marco Körner

Technical University of Munich (TUM), TUM School of Engineering and Design,
Department of Aerospace and Geodesy, Arcisstr. 21, 80333 Munich, Germany

{maja.schneider, marco.koerner}@tum.de

ABSTRACT

With leaps in machine learning techniques and their application on Earth observation challenges has unlocked unprecedented performance across the domain. While the further development of these methods was previously limited by the availability and volume of sensor data and computing resources, the lack of adequate reference data is now constituting new bottlenecks. Since creating such ground-truth information is an expensive and error-prone task, new ways must be devised to source reliable, high-quality reference data on large scales. As an example, we showcase EUROCRIPS, a reference dataset for crop type classification that aggregates and harmonizes administrative data surveyed in different countries with the goal of transnational interoperability.

Index Terms— administrative data, crop classification, machine learning, reference data, ground-truth

1. INTRODUCTION

In recent years, data-driven methods addressing remote sensing and Earth observation problems have shown impressive performance [3]. While the development of such approaches used to be limited by the amount of available observational data and computing resources, these determining factors have now vanished. Modern Earth observation programs provide a multitude of data products across manifold spectral, spatial, and temporal resolutions with today's data processing pipelines able to process these data volumes [17]. These have been used to compile various general-purpose [12] or application-specific datasets [11, 18]. Instead, the lack of appropriate reference data—*i.e.*, labels, annotations, or targets—is now the new bottleneck that restricts the further development of data-driven modelling and information extraction techniques. To keep

pace with the ever-growing and expanding data archives, the properties of reference data must align with those of the Earth observation data. This, in particular, calls for improved quantity, quality, resolution, and temporal frequency of reference data. However, this is difficult to achieve with established annotation processes, *e.g.*, manual labelling procedures or iterative and interactive curation protocols.

To address this problem, we want to motivate the use of *administrative data* and describe its far-reaching possibilities. With the example of the EUROCRIPS initiative, we demonstrate how pre-existing metadata can be used to derive reliable, high-quality and interoperable reference datasets on large spatial and temporal scales.

2. ADMINISTRATIVE DATA

With an evergrowing focus on data production and collection across society, the volume of *administrative data* or *government data* has also been increasing exponentially in recent years [8]. This data is acquired, collected, or compiled by public authorities to support and enable government services and processes, including planning or monitoring, and thus serves as the basis for decisions and interventions.

To serve these purposes, administrative data is usually collected in a continuous and regular manner. These collections are characterised by high granularity and wide coverage, must meet strict quality as well as reliability standards, and need to be well documented. Administrative data may contain items that provide unique keys across different sources, *e.g.*, geocodes, instance identifiers, acronyms or pseudonyms, or further properties.

Despite all of this, the use of administrative data is complex and difficult. As a consequence of their distributed acquisition, data is usually available in different, often proprietary, formats or lacking any standardisation, limiting its use and interoperability. Likewise, incompatible formats require prior harmonisation, alignment, and aggregation or disaggregation [4]. The fact that administrative data is surveyed from different contexts and organisational levels—*e.g.*, on supra-national, trans-/international, national, sub-national, (inter-)regional, or

*corresponding author

M. Schneider, M. Körner, and EUROCRIPS receive funding from the German Federal Ministry for Economic Affairs and Climate Action on the basis of a resolution of the German Bundestag under reference 50EE1908 and from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101004112. The authors thank the Stifterverband for supporting EUROCRIPS with the Open Data Impact Award 2021.

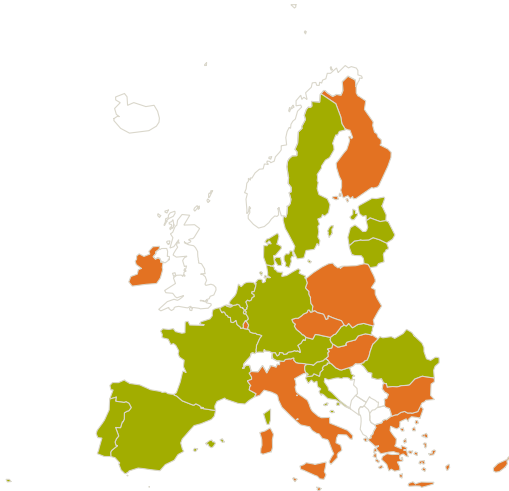


Fig. 1: EUROCROPS compiles crop type reference data from several *European Union (EU)* countries. This data has been extracted from self-declarations that farmers submit to receive subsidies under the *common agricultural policy (CAP)*. Several countries, shaded in *green*, already publish this information, in contrast to those shaded in *orange*.

at community level—constitutes another impediment to its further use. Finally, there is a lack of central and global repositories, obscuring the existence of datasets to scientists and engineers, while their use is further restricted by privacy protection and data agency legislation, their closed-source nature and their huge volumes of data.

Considering the aforementioned observations, the use of administrative data remains to offer enormous potential [2]. Being able to access and derive datasets from them allows for cross-validation to assess and improve the quality of the data. The wide coverage of continuously maintained administrative data enables both spatial and temporal analysis to be conducted. Encouragingly, ever more countries are developing their open data strategies [7], foreshadowing that administrative data will be widely and affordably available in the not too distant future.

3. THE EUROCROPS PROJECT

These considerations discussed so far speak of the difficulties of using administrative data for scientific purposes. To counter this, we initiated the EUROCROPS project, where we aggregate and harmonise administrative data surveyed in different countries of the *European Union (EU)*, with the goal of transnational interoperability.



Fig. 2: EUROCROPS compiles precise geometric representations of each field parcel and enriches these with their associated *hierarchical crop and agriculture taxonomy (HCAT)*-encoded crop class label.

3.1. Motivation and Idea

Instantiating a motivating use-case, we have chosen the problem of crop type classification from optical remote sensing imagery. This is an inherently difficult task, as vegetation features remarkable biological and geographic diversity and, therefore, requires high-capacity approaches to be modelled in a data-driven way. Further, vegetation processes are intrinsically time-dependent, *i.e.*, temporal sequences of observations are required to capture the underlying dynamics appropriately. Fortunately, modern Earth observation satellites provide this information in abundance, and current data-driven machine learning approaches are showing unprecedented performance in predicting and classifying plant species [9, 10, 15, 18]. Nevertheless, these approaches are limited by the amount and quality of the corresponding ground-truth information.

3.2. Database

Manually annotating Earth observation data with the particular cultivated plant species grown on each field parcel (*cf.* Fig. 2) becomes infeasible on large scales; new ways of acquiring this information are in need. For EUROCROPS, we take advantage of the *common agricultural policy (CAP)* set up by the EU. This requires the farmers of each member state to declare the crop species cultivated on their parcels in order to receive subsidies accordingly. Although data is collected and archived by the authorities of the respective state, the majority of them keep it undisclosed.

EUROCROPS aims at compiling and harmonising such data to demonstrate its widespread potential.

3.3. Taxonomy

EUROCROPS is intended to be used in combination with any kind of remote sensing data and transnationally across spatial scales. As the various country-specific protocols to represent

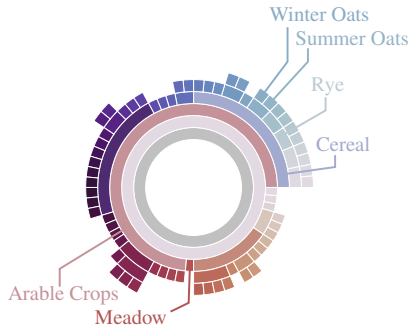


Fig. 3: The *hierarchical crop and agriculture taxonomy (HCAT)* [13] uniquely encodes each crop species and allows to compare self-declaration data transnationally across countries. Note that this visualization has been simplified for visualization purposes.

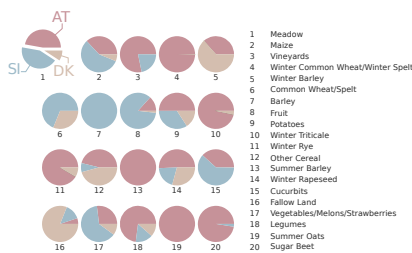


Fig. 4: EUROCRIPS allows for analysing the statistics of cultivated crops and their comparison across the different contributing countries. This information can be used to derive biodiversity indicators.

crop species are typically mutually incompatible, it was necessary to harmonise them appropriately. This gave rise to the HCAT [13], extending the existing *EAGLE* matrix [1]. As visualized in Fig. 3, each crop species can be uniquely encoded with a specific multi-level HCAT identifier

$$\begin{array}{ccccccc} \boxed{33} & - & \boxed{XX} & - & \boxed{XX} & - & \boxed{XX} & - & \boxed{XX} \\ \text{position in} & & \text{Level 3} & & \text{Level 4} & & \text{Level 5} & & \text{Level 6} \\ \text{EAGLE matrix} & & & & & & & & \end{array}$$

ranging from *Cereals* as 33-01-01-00-00 (level 4) to *Summer Oats* as 33-01-01-05-03 (level 6). This representation scheme allows for comparing crop species across contributing EU countries.

3.4. Fields of Application

EUROCRIPS is data-agnostic and designed to be used together with any kind of geo-referenced Earth observation data. Thus,

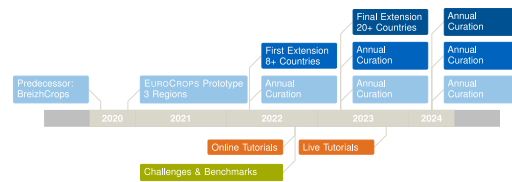


Fig. 5: The work on EUROCRIPS project is scheduled for the next years until 2024. Beyond the primary data collection activities carried out continuously during the entire project time span, we plan to use EUROCRIPS as a solid basis for creating teaching and learning concepts and to build reliable benchmarking tools that will be offered to the public.

the field of possible applications is vast and ample. By addressing the motivating use-case of crop type classification using this reference dataset, data-driven methods will implicitly learn internal representations of vegetation dynamics, *i.e.*, its *phenology*. Such models can further be used to monitor the development of vegetation stocks to spot unexpected patterns, *e.g.*, caused by environmental influences [6], or to predict the expected yield [5]. Furthermore, the country-specific statistics compiled and harmonised in EUROCRIPS allow for statistical investigation of regional distributions of crop cultivation patterns (*cf.* Fig. 4) and to derive biodiversity markers at various regional scales [14]. Owing to its broad applicability, EUROCRIPS provides the basis for further research projects.¹

3.5. Current State

Since the start of the project, we collected declaration data from 16 countries across the years 2015 to 2021 (*cf.* Fig. 1) and harmonised one year per country. To demonstrate the use of EUROCRIPS, we compiled the early demonstrator dataset TINYEUROCRIPS [16] containing reference data from *Austria*, *Germany*, and *Slovenia*. The current state of EUROCRIPS can be tracked through the project website www.eurocrops.tum.de and the associated GITHUB repository github.com/maja601/EuroCrops. The latter also serves as a platform for tutorials, ongoing discussions, requests, and bug tracking.

3.6. Plans and Outlook

With EUROCRIPS, we not only want to provide the reference dataset described hitherto but rather to demonstrate its far-reaching potential on a broad scale. As visualised in Fig. 5, the project envisages several further initiatives:

Annual Curation and Extension The EUROCRIPS project schedule foresees actions until the year 2024. Within that

¹*e.g.*, *Global Earth Monitor* (www.globalearthmonitor.eu), *Pre-TrainAppEO* (www.asg.ed.tum.de/lmf/pretrainappec), *DUKE* (www.asg.ed.tum.de/lmf/duke), *etc.*

time, continuous updates and temporal extensions to the existing dataset will be applied with new reference data. Simultaneously, more countries will be added, as they volunteer to contribute their data.

Tutorials We intend to use the EURO-CROPS dataset as a sound foundation to introduce further user groups to the emerging topics of machine learning in the context of Earth observation. For this purpose, teaching and learning modules for self-study, as well as for the use in interactive tutorial formats, will be developed. These workshops will be held at upcoming community events, conferences and workshops.

Benchmarks and Challenges In order to assist researchers and users in the evaluation of their data processing pipelines, we will further provide tools and methods to benchmark several tasks related to crop type classification using regionally and temporally held-out subsets of EURO-CROPS.

4. SUMMARY

The increase of remote sensing data and compute resources within recent years no longer restrict the development of data-driven Earth observations models. Instead, the community is now lacking reference data matching the properties of the data in remote sensing archives. To counteract this new bottleneck, we presented EURO-CROPS to showcase how existing administrative data can be used to derive high-quality, reliable ground-truth annotations to train and evaluate data-driven remote sensing and Earth observation models. With its analysis-ready design and widespread availability, EURO-CROPS also targets experts from other research communities and invites them to address prevalent problems across remote sensing analysis and Earth observation research.

EURO-CROPS will be made available through several remote sensing data repositories, e.g., GEO-DB, CODE-DE, EO-LEARN, and GOOGLE EARTH ENGINE.

5. REFERENCES


- [1] S. Arnold, B. Kosztra, G. Banko, G. Smith, G. Hazeu, M. Bock, and N. V. Sanz, "The EAGLE concept, A vision of a future european land monitoring framework," in *EARSeL Symposium proceedings, "Towards Horizon 2020"*, 2013, pp. 551–568.
- [2] R. Connelly, C. J. Playford, V. Gayle, and C. Dibben, "The role of administrative data in the big data revolution in social science research," *Social Science Research*, vol. 59, pp. 1–12, 2016. DOI: [10.1016/j.ssresearch.2016.04.015](https://doi.org/10.1016/j.ssresearch.2016.04.015).
- [3] O. Dubovik, G. L. Schuster, F. Xu, Y. Hu, H. Bösch, J. Landgraf, and Z. Li, "Grand challenges in satellite remote sensing," *Frontiers in Remote Sensing*, vol. 2, 2021. DOI: [10.3389/frsen.2021.619818](https://doi.org/10.3389/frsen.2021.619818).
- [4] C. Marini and V. Nicolardi, "Big data and economic analysis: The challenge of a harmonized database," in *Studies in Classification, Data Analysis, and Knowledge Organization*, Springer, 2020, pp. 235–246. DOI: [10.1007/978-3-030-51222-4_18](https://doi.org/10.1007/978-3-030-51222-4_18).
- [5] M. Marszalek, M. Körner, and U. Schmidhalter, "Prediction of multi-year winter wheat yields at the field level with satellite and climatological data," *Computers and Electronics in Agriculture*, vol. 194, p. 106777, 2022. DOI: [10.1016/j.compag.2022.106777](https://doi.org/10.1016/j.compag.2022.106777).
- [6] M. Marszalek, M. Lösch, M. Körner, and U. Schmidhalter, "Early crop-type mapping under climate anomalies," *Preprints*, 2022. DOI: [10.20944/preprints202004.0316.v2](https://doi.org/10.20944/preprints202004.0316.v2).
- [7] A. Quarati, "Open government data: Usage trends and metadata quality," *Journal of Information Science*, pp. 1–24, 2021. DOI: [10.1177/01655515211027775](https://doi.org/10.1177/01655515211027775).
- [8] D. Reinsel, J. Rydning, and J. F. Gantz, "Worldwide global data-sphere forecast, 2021–2025: The world keeps creating more data. Now, what do we do with it all?" International Data Corporation (IDC), research rep. US46410421, 2021.
- [9] M. Rußwurm and M. Körner, "Multi-Temporal Land Cover Classification with Sequential Recurrent Encoders," *ISPRS International Journal of Geo-Information*, vol. 7, no. 4, p. 129, 2018. DOI: [10.3390/ijgi7040129](https://doi.org/10.3390/ijgi7040129).
- [10] 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. DOI: [10.1016/j.isprsjprs.2020.06.006](https://doi.org/10.1016/j.isprsjprs.2020.06.006).
- [11] M. Rußwurm, C. Pelletier, M. Zollner, S. Lefèvre, and M. Körner, "BreizhCrops: A time series dataset for crop type mapping," in *ISPRS – International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. XLIII-B2-2020, 2020, pp. 1545–1551. DOI: [10.5194/isprs-archives-XLIII-B2-2020-1545-2020](https://doi.org/10.5194/isprs-archives-XLIII-B2-2020-1545-2020).
- [12] M. Schmitt, L. H. Hughes, C. Qiu, and X. X. Zhu, "SEN12MS – a curated dataset of georeferenced multi-spectral sentinel-1/2 imagery for deep learning and data fusion," *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. IV-2/W7, pp. 153–160, 2019. DOI: [10.5194/isprs-annals-iv-2-w7-153-2019](https://doi.org/10.5194/isprs-annals-iv-2-w7-153-2019).
- [13] M. Schneider, A. Broszeit, and M. Körner, "EuroCrops: A pan-european dataset for time series crop type classification," in *Proceedings of the Conference on Big Data from Space (BiDS)*, Publications Office of the European Union, 2021. DOI: [10.2760/125905](https://doi.org/10.2760/125905).
- [14] M. Schneider, D. Gackstetter, S. T. Meyer, T. Schelte, F. Schmitz, and M. Körner, "Analysing the impact of european agriculture on biodiversity with an updated hierarchical crop and agriculture taxonomy," *npj Biodiversity*, in preparation.
- [15] M. Schneider and M. Körner, "[re] satellite image time series classification with pixel-set encoders and temporal self-attention," *ReScience C*, vol. 7, 2 2021. DOI: [10.5281/zenodo.4835356](https://doi.org/10.5281/zenodo.4835356).
- [16] M. Schneider and M. Körner, *TinyEuroCrops*, Dataset, Technical University of Munich (TUM), 2021. DOI: [10.14459/2021MP1615987](https://doi.org/10.14459/2021MP1615987).
- [17] M. Sudmanns, D. Tiede, S. Lang, H. Bergstedt, G. Trost, H. Augustin, A. Baraldi, and T. Blaschke, "Big earth data: Disruptive changes in earth observation data management and analysis?" *International Journal of Digital Earth*, vol. 13, no. 7, pp. 832–850, 2020. DOI: [10.1080/17538947.2019.1585976](https://doi.org/10.1080/17538947.2019.1585976).
- [18] M. O. Turkoglu, S. D'Arconco, G. Perich, F. Liebisch, C. Streit, K. Schindler, and J. D. Wegner, "Crop mapping from image time series: Deep learning with multi-scale label hierarchies," *Remote Sensing of Environment*, vol. 264, p. 112603, 2021. DOI: [10.1016/j.rse.2021.112603](https://doi.org/10.1016/j.rse.2021.112603).

3 [Re] Satellite Image Time Series Classification with Pixel-Set Encoders and Temporal Self-Attention

Maja Schneider and Marco Körner (2021a). '[Re] Satellite Image Time Series Classification with Pixel-Set Encoders and Temporal Self-Attention'. In: *ReScience C* 7.2, #19. DOI: [10.5281/zenodo.4835356](https://doi.org/10.5281/zenodo.4835356)

Replication / ML Reproducibility Challenge 2020

[Re] Satellite Image Time Series Classification with Pixel-Set Encoders and Temporal Self-Attention

Maja Schneider^{1, } and Marco Körner^{1, }¹Chair of Remote Sensing Technology, Department of Aerospace and Geodesy, Technical University of Munich (TUM), Munich, Germany

Edited by
Koustuv Sinha,
Jesse Dodge

Reviewed by
Anonymous Reviewers

Received
29 January 2021

Published
27 May 2021

DOI
10.5281/zenodo.4835356

Reproducibility Summary

The presented study evaluates “Satellite Image Time Series Classification with Pixel-Set Encoders and Temporal Self-Attention” by Garnot et al. [1] within the scope of the ML Reproducibility Challenge 2020. Our work focuses on both aspects constituting the paper: the method itself and the validity of the stated results. We show that, despite some unforeseen design choices, the investigated method is coherent in itself and performs the expected way.

Scope of Reproducibility

The evaluated paper presents a method to classify crop types from multispectral satellite image time series with a newly developed *pixel-set encoder* and an adaption of the Transformer [2], called *temporal attention encoder*.

Methodology

In order to assess both the architecture and the performance of the approach, we first attempted to implement the method from scratch, followed by a study of the authors’ openly provided code. Additionally, we also compiled an alternative dataset similar to the one presented in the paper and evaluated the methodology on it.

Results

During the study, we were not able to reproduce the method due to a conceptual misinterpretation of ours regarding the authors’ adaption of the Transformer [2]. However, the publicly available implementation helped us answering our questions and proved its validity during our experiments on different datasets. Additionally, we compared the papers’ temporal attention encoder to our adaption of it, which we came across while we were trying to reimplement and grasp the authors’ ideas.

What was easy

Running the provided code and obtaining the presented dataset turned out to be easily possible. Even adapting the method to our own ideas did not cause issues, due to a well documented and clear implementation.

Copyright © 2021 M. Schneider and M. Körner, released under a Creative Commons Attribution 4.0 International license.

Correspondence should be addressed to Maja Schneider (maja.schneider@tum.de)

The authors have declared that no competing interests exist.

Code is available at <https://github.com/maja601/pytorch-psetae>. – SWH swh:1.dir:631305811058b775406ae624095e3ec66ace6437.

Data is available at <http://www.eurocrops.tum.de/>.

[Re] Satellite Image Time Series Classification with Pixel-Set Encoders and Temporal Self-Attention

What was difficult

Reimplementing the approach from scratch turned out to be harder than expected, especially because we had a certain type of architecture in mind that did not fit the dimensions of the layers mentioned in the paper. Furthermore, knowing how the dataset was exactly assembled would have been beneficial for us, as we tried to retrace these steps, and therefore would have made the results on our dataset easier to compare to the ones from the paper.

Communication with original authors

While working on the challenge, we stood in E-mail contact with the first and second author, had two online meetings and got feedback to our implementation on GITHUB. Additionally, one of the authors of the Transformer paper [2] provided us with further answers regarding their models' architecture.

1 Introduction

The *machine learning* community showcases impressively and with great success how to design systems for analysing large data inventories, with the goal of identifying relevant patterns within them. At the same time, the *remote sensing* and *Earth observation* sector found itself confronted with the availability of novel dedicated sensor platforms. These are now capable of continuously acquiring new data at high temporal, spatial, and spectral frequencies, requiring the development of innovative and efficient ways to process these data stocks. In light of that, several machine learning methods have found wide application in the field of remote sensing and Earth observation. As the availability of observation data increased massively during the last decades [3], various retrieval, detection, and prediction problems can be addressed this way. Nevertheless, Earth observation data is mostly of a very inhomogeneous nature, which is due to the different designs and layouts of the receiving sensor platforms. In addition, geophysical processes on the Earth's surface are complex and manifest themselves in observable changes with different dynamic patterns. Therefore, the goal of gaining deeper understanding of such processes requires the use of interpretable models [4].

In their recent CVPR publication "Satellite Image Time Series Classification with Pixel-Set Encoders and Temporal Self-Attention," Garnot *et al.* [1] propose a new method to address these issues. Motivated by the practical problem of *crop type classification* from sequences of optical satellite imagery—that we consider a proxy for the entirety of vegetative processes on the Earth's surface—the authors made use of *attention mechanisms*. In particular, they claimed adapting the *Transformer* architecture [2] that has gained considerable popularity in the recent past, enabling it to digest such specific Earth observation data modalities. Additionally, they introduce a *pixel-set encoder* as a new option to deal with medium-resolution satellite images instead of the known convolutional neural networks in image processing.

Due to their aforementioned properties, the handling of Earth observation data in practical applications requires special care. Most prevalently, they exhibit a considerable amount of spatial autocorrelation [5], reinforcing the already known issues of *underspecification* inherent to data-driven machine learning models [6]. This generally leads to overfitting effects and poor generalisation performance. Hence, we carefully reproduced the proposed methods, first by starting from scratch just following the descriptions given in the paper under investigation, and subsequently by adapting the reference implementation openly provided by the original authors. To sanity-check both implementations and to further assess the transferability and generalisation properties of the studied model, we carried out further experiments relying on an alternative but comparable dataset. Following the idea of this reproducibility challenge, we will make our implementation and data public, allowing the community to likewise evaluate our findings.

1.1 Reproducibility questions

As a foundation of our reproduction study, we identified the following key questions, each one examining one particular aspect or claim of the original paper:

- i) Is it possible to reproduce the presented methods and their performance with *and* without referring to the authors' publicly available code?
- ii) To which extent do the author's implicit claim of adapting the transformer architecture affect the model and its performance?
- iii) Does the model perform comparably well when being only tested or both trained and tested on a different dataset?

[Re] Satellite Image Time Series Classification with Pixel-Set Encoders and Temporal Self-Attention

- iv) How is the outcome influenced by the choice of the test set assembly, namely by splitting the data randomly or regionally?

1.2 Contributions beyond the original paper

In addition to the results in the original paper, we also report the performance of the method evaluated on an alternative dataset and on a test set that does not show regionally overlap with the training set.

2 Methodology and experimental setup of the reproduction study

In order to study the reproducibility of the original publication, we followed different approaches to answer the questions raised in Section 1.1. Each of the following subsections will address one of these questions by introducing the methodology behind the chosen experiments and present our obtained results as replies to the author's claims. Generally, we traversed the following three different experimental stages that we will also publish on GITHUB¹ for better traceability: We first started by developing the entire approach ourselves in PYTHON and PYTORCH [7], as described in Section 2.1.1, followed by a short study on the original implementation in Section 2.1.2 and a comparison of one particular technical aspect in Section 2.2. Eventually, we conducted experiments on the influence of input data in the remaining sections.

2.1 Reproduction and accuracy of the satellite time series classification

The proposed architecture for *satellite time series classification* consists of two components which the authors introduce as the *pixel set encoder (PSE)* and the *temporal attention encoder (TAE)*. While the former takes care of a randomly sampled pixel set from a crop parcel and produces an embedding of the input, the latter, an adapted variant of the Transformer architecture [2], produces an output by applying self-attention to these multi-temporal embeddings. Unlike practices familiar from *natural language processing*, PSE and TAE get both optimised during the training phase. A detailed summary of the composition and number of parameters in the networks can be obtained from Table 1 in the studied paper.

All experiments were conducted on an UBUNTU 20.10 workstation equipped with 64 GB of RAM, an INTEL I7-8700 CPU and an NVIDIA GEFORCE RTX 2060 GPU.

Full replication study of the approach – We reimplemented the proposed architecture described in Section 3: *Methods* of the investigated paper almost literally in PYTHON. As the section is subdivided into the three parts describing the *spatial encoder*, also referred to as the *pixel-set encoder*, the temporal attention encoder, and the *spatio-temporal classifier*, these three modules likewise build the core of this reproduction study. More precisely, Table 1 of Garnot *et al.* [1] allowed us to inherit all model hyper-parameters straightforwardly. For one of the four used *multi-layer perceptrons (MLPs)*, it was stated that it consisted of *fully-connected layers FC*, *batch normalisation*, and *ReLU activations*. We, thus, inductively assumed these components to be part of the other three MLPs as well. While we managed to develop an inefficient yet working version of the spatial encoder, some aspects of the *temporal attention encoder* appeared unintuitive at first sight. Unlike the original Transformer model [2], on which this module is based, the values v are not calculated by a fully-connected layer. Instead, the sum of the *spatial encoder* outputs and the positional encoding is multiplied with the attention mask a , which is visualised in Figure 1a. The authors motivated this change with the claim that it “removes needless computations, and avoids a potential information bottleneck” [cf. 1, Section

¹<https://github.com/maja601/RC2020-psetae>

3.2.]. Having the original Transformer implementation in mind, we misinterpreted this design choice. Thus, in our implementation, we divided the multi-head attention input into four equal tracks, as we were not able to think of another way to end up having 512 nodes to be passed over to the MLP_3 (cf. Table 1 of [1]). This misconception was partially reinforced by the superscript (h) in formula (5) in [1], i.e.,

$$k_h^{(t)}, q_h^{(t)} = \text{FC}_1^{(h)}(e^{(t)} + p^{(t)}) \quad , \quad (1)$$

suggesting that, for each head h , an entirely independent fully-connected layer $\text{FC}^{(h)}$ was used. This way, our network reimplementaion became incredibly blown up and we were not able to spot the correct approach.

All hyper-parameters and training details were directly taken from the Section 3.4: *Implementation details* from Garnot *et al.* [1]. After completing the first presented stages of the implementation, we were able to achieve an accuracy of about 60 %. Subsequently, we had a first online meeting with the first and second author of the investigated paper, where we identified some misconception concerning the used labels: Instead of using all 20 classes from the `label_19class` dictionary in the provided `lables.json` file, only the top-20 classes from the `label_44class` dictionary of the same file were utilised by the authors, i.e., the classes with more than 100 occurrences in the dataset. It also got to our attention that, *in lieu* of the proposed batch normalisation, the multi-layer perceptrons should perform *layer normalisation*. Unfortunately, despite the first author providing helpful feedback on our implementation via GITHUB, we were still not able to achieve a relevant increase in accuracy compared to the previously stated 60 %. After comparing our implementation to the reference implementation provided by the authors, it became apparent that we struggled to grasp the authors' ideas about the data organisation and the spatio-temporal classifier. Therefore, we will investigate what led us to misinterpret the Transformer's adaption in Section 2.2 and the data organisation in Section 2.3.2.

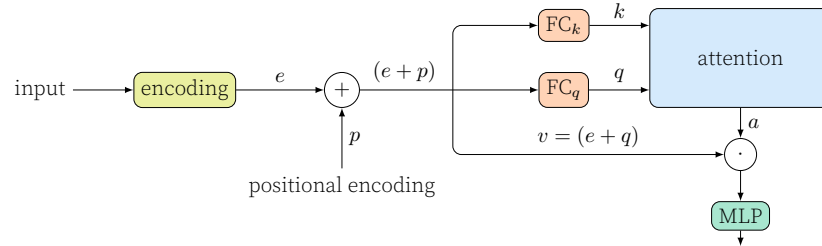
Evaluation of the original implementation – Obtaining the authors' code and running it locally proved to be easy thanks to a well-documented GITHUB repository². In general, their reference implementation is modular, clearly structured, and sufficiently commented which makes the entire architecture easy to adapt to one's own needs. The model can either be trained from scratch or, together with a provided checkpoint, used to solely run the inference. We present a comparison of the test results from the checkpoint and when training from scratch in Table 1a. Using the system specified in Section 2.1, training the model for one single epoch took approximately 27 s, while running the inference on one sample batch containing 128 parcels completed within about 0.04 s. For the following experiments, we exclusively used the official reference implementation to ensure comparability.

2.2 Experiments on the transformer architecture

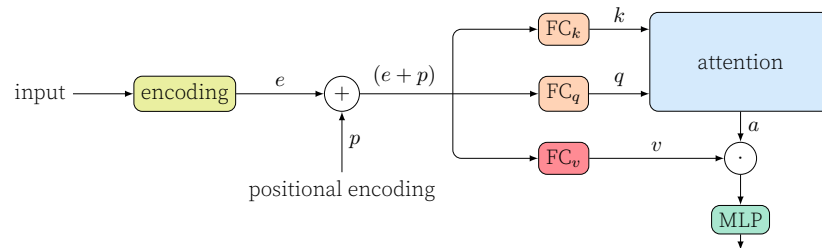
As stated in Section 2.1.1, we previously relied on some faulty assumptions related to the architecture of the adapted Transformer multi-head attention. Fortunately, the original code provided by the authors helped us to reconstruct their initial idea. Nevertheless, having the vanilla off-the-shelf Transformer implementation in mind, we were inquisitive about the impact of the particular architectural change proposed by Garnot *et al.* Hence, we took the reference implementation and changed two lines of it in a way that the temporal attention encoder then employed a third fully-connected layer FC_v , just like the vanilla Transformer attention model does, as described by Vaswani *et al.* [2]. Figure 1b illustrates this subtle change in comparison to the architecture realized by

²<https://github.com/VSainteuf/pytorch-psetae>

[Re] Satellite Image Time Series Classification with Pixel-Set Encoders and Temporal Self-Attention



(a) The temporal attention encoder according to the presented paper, where the values v are not calculated specifically, but the sum of e and p is directly passed through to the dot product.



(b) Our adaption of the temporal attention encoder with the additional fully-connected layer FC_v , as mentioned in “Attention is All you Need” by Vaswani *et al.* [2].

Figure 1. Illustration of the architectural difference between (a) the proposed temporal attention encoder and (b) the extended version of it presented in this study. While Garnot *et al.* approached satellite image time series classification with a lightweight adaption of the Transformer architecture [2], we kept the original third fully-connected layer FC_v . The additional building blocks of the surveyed method—namely the encoder, positional encoding, attention, and final MLP—remained unchanged. This schematic representation does not include the multi-head attention, which mainly influences the *attention* module and would increase the complexity unnecessarily.

the authors of the paper under investigation, shown in Figure 1a. We provide the code for this version as a fork of the original repository.³

While this adaption of the proposed architecture reduces the number of trainable model parameters to 80%, we were able to reproduce the reported performance or even experienced a slight increase in accuracy, as summarised in Table 1a.

2.3 Generalisation and transferability

Until that stage, we mainly considered the theoretical aspects of this reproduction study. Conversely, this section focuses on the application-oriented side. In light of our observations described before, we wanted to investigate whether the increased number of parameters in the original model can become beneficial when scaling the underlying classification problem by confronting it with an alternative dataset.

Datasets – Although a tremendous amount of satellite data, especially from the SENTINEL-2 platforms, is publicly available, the computer vision and machine learning community still lacks labels or annotations for addressing most relevant research questions. Therefore, Garnot *et al.* [1] did not only publish a new method, but also complemented it with a dataset containing crop type labels of more than 190 000 agricultural parcels within the area of a particular tile of the SENTINEL-2 tiling grid T31TFM, located in France.

³<https://github.com/maja601/pytorch-psetae>

Table 1. Covering several experiments, (a) and (b) show the results of the presented temporal attention encoder (TAE) and our adaptation of it with the additional fully-connected layer FC_v , respectively. We therefore hoped to give an impression of the performance of the models under different circumstances and also the impact of the architectural change we evaluated in Section 2.2.

(a) Comparison of the original TAE and its adaption with FC_v . The overall accuracy on FR-T31TFM for the approach of Garnot *et al.* is given twice: once it was obtained from the checkpoint provided by the authors and once by training from scratch. The last column indicates how many trainable parameters each of the network variants comprises.

| Model variant | Overall accuracy | | Number of parameters |
|---------------------|------------------|--------------|----------------------|
| | checkpoint | trained | |
| TAE [1] | 94.19 | 94.26 | 164 116 |
| TAE with FC_v [2] | — | 94.24 | 131 476 |

(b) Analogous to (a), both versions of the models were trained on FR-T31TFM, but this time tested on the unseen region SI034 of the SI-T33TWM dataset. This practice is often referred to as cross-dataset evaluation.

| Model variant | Overall accuracy | |
|---------------------|------------------|--------------|
| | checkpoint | trained |
| TAE [1] | 61.75 | 61.65 |
| TAE with FC_v [2] | | 62.03 |

Additionally and to evaluate the reproducibility of the presented methods' results on a different input, an analogous dataset with parcels located in Slovenia was constructed in the course of this study. The following two sections gives insight into the background and properties of these two datasets.

FR-T31TFM: Dataset from the paper Unlike CNN-based methods, the approach by Garnot *et al.* does not require the observation data to be stored as images with defined neighborhood relations, but rather as an unordered set of pixels for each parcel. From their GITHUB page, a toy dataset containing 500 parcels, each saved as NUMPY data files, can be obtained. To get access to the entire dataset, an inquiry needs to be sent to the authors, which they reply to within no time. This dataset includes 192 056 NUMPY data files of dimension $T \times C \times N$, with $T, C = 10$, and N being the number of observations dates, spectral bands, and pixels for each particular parcel, respectively. It additionally comes with several metafiles, *i.e.*, the dates of the observations, the labels of the parcels, the geometric features of the parcels—which are stated to be necessary for the pixel set encoder—and pre-computed normalisation values. By design, the dataset is randomly partitioned into test, validation, and training parcels, following a split ratio of 3:1:1. This dataset also faces one of the biggest challenges in crop type classification from satellite data, namely the uneven distributed data foundation, as visualised in Figure 3a. In their original paper, Garnot *et al.* refer to several data preprocessing steps, such as reducing the number of spectral bands delivered by the SENTINEL-2 satellite from 13 to 10, linearly interpolating ground pixels that are affected by cloud cover, normalising the reflectance data, and adding Gaussian noise.

SI-T33TWM: Additional dataset As part of another project at our lab, a pan-European reference dataset for crop type classification is currently under development and will be made publicly available early next year. This way, we had the chance to use some of the data obtained from Slovenia to construct a pixel set similar to the one presented by Garnot *et al.* [1]. In order to keep it as close as possible to the original, we selected similar time steps and also performed a linear cloud pixel interpolation on L2A SENTINEL-2 data. As we did not have access to the precise cloud detection module used by the authors, we manually annotated cloudy pixels in this region of interest. The `dataset_preparation.py` script provided by Garnot *et al.* took care of pooling the Sentinel-2 data and the reference data in GEOJSON format. The necessary normalisation parameters had likewise to be calculated manually, as well as the operation to extract the geometric features used in the spatial encoder. Contrasting the description given in the paper under investigation, the sequential order of components within this geometry feature vector differed in a way

[Re] Satellite Image Time Series Classification with Pixel-Set Encoders and Temporal Self-Attention

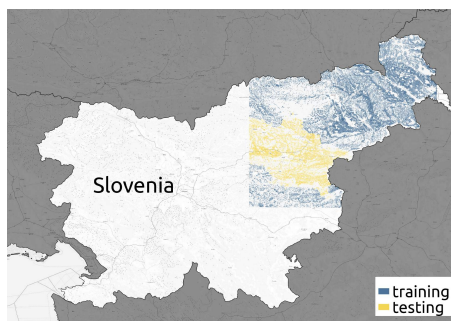


Figure 2. Schematic illustration of the additional SI-T33TWM dataset located in Slovenia with reference crop parcels mainly covering the north-eastern part of the country. While the blue areas are used to evaluate the train and test procedure proposed by Garnot *et al.*, the yellow area was set aside in advance. This way, we ensured to have an extra regionally differentiated test set that does not suffer from the issue of spatial autocorrelation.

that its second element appeared not to be the *pixel count* N but rather, as we assume from looking at the original dataset, the area of the bounding box. Considering the labels, we prepared two versions: One file with the previously stated top-20 crop classes from the original dataset, called top-20-F, and another one containing the top-20 classes from our alternatively chosen region, analogously named top-20-S and illustrated in Figure 3b. Since the crop cultivation in Slovenia differs from France, several classes of the original top-20-F were not represented in the new dataset, as shown in Figure 3c. This appears to render the classification on top-20-F an easier problem than on top-20-S. Section 2.4 will provide more background regarding the split of the dataset and to one of the particular difficulties inherent to geospatial data, which is why we put one region aside in advance. This way, we were able to evaluate the performance of the method also on regionally unseen data. A visual explanation of the train and test split is shown in Figure 2.

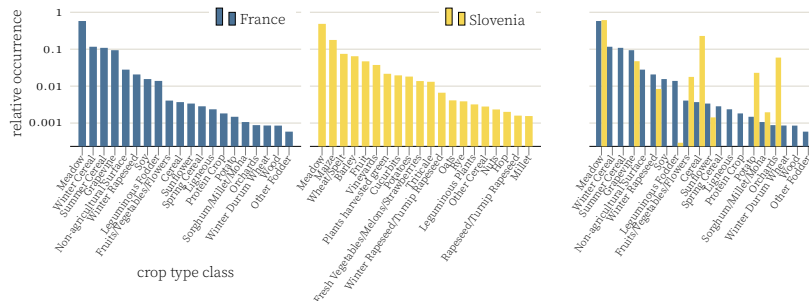
Cross-dataset evaluation – The first experiment on the generalisation abilities of the described method was performed with the model being trained on the FR-T31TFM dataset. After the descriptions of the chosen classes were obtained from the original authors, we picked the same classes from our SI-T33TWM dataset and ran the inference on our prepared test pixel set. Table 1b summarises the overall results, showing that the performance dropped significantly compared to the ones reported previously (*cf.* Table 1a).

Method application to a different region – Taking advantage of the relatively short training time, it was easily possible to train the entire model on our own data. This way, we evaluated whether the outcomes reported by Garnot *et al.* could be reproduced, even without having access to data at exactly the same preprocessing level. For this purpose, we ran the experiment twice on the new SI-T33TWM dataset, *i.e.*, first with the top-20-F labels and then with the top-20-S labels. In Table 2, the columns *random split* show the results of these experiments. These can directly be compared to the ones from the paper.

2.4 Experiments on the choice of the test set

Beside the issue of limited geometric resolution in SENTINEL-2 data, the influence of spatial autocorrelation has always to be taken into account when dealing with Earth obser-

[Re] Satellite Image Time Series Classification with Pixel-Set Encoders and Temporal Self-Attention



(a) The relative class occurrences of the crop types in the FR-T31TFM dataset provided by the authors of the reproduced paper. (b) The class distribution of the 20 most frequent classes of the newly constructed SI-T33TWM dataset. (c) A comparison of the class distribution and the relative occurrences of these classes (a) in the new SI-T33TWM dataset. This shows that training and testing on the SI-T33TWM while using the top-20-F labels will lead to fewer classes in total.

Figure 3. Statistics on the different datasets used in this study. As the data harmonisation was done by hand, the class names of one dataset might include differently named crop types from the other dataset and vice versa. We point out that the relative occurrence is indicated in log-scale, highlighting the strong class imbalance towards meadow.

vation satellite imagery [5]. Due to the coarse sampling of the Earth’s surface, adjacent pixels might share relevant amounts of information with each other. Therefore, using contiguous field parcels for training as well as for testing can lead to non-representative results over-estimating the true performance of the classification model. We tried to account for these effects by reserving an entirely separate region, shown in Figure 2, as an additional test set. This region was selected to approximate one-fifth of the parcels and a class distribution representative for the entire SI-T33TWM dataset. Results of this experiment are included into Table 2, where we compare the original random split to the new regional split.

3 Discussion

This section focuses on the analysis of reproducibility in general and will not justify the authors claim that their method is the current state-of-the-art approach to solve satellite time series classification. We therefore split the findings of the study into two aspects that we tried to evaluate: On the one hand, we will discuss the insights we have gained by reproducing the methodological process itself, and on the other hand, we will elaborate our approach to reproduce the desired results produced by the method.

3.1 Reproducibility of the method

When reimplementing the full architecture, we found that, despite having all parameters at hand, it would have been helpful to have access to more information concerning the data preprocessing and organisation, as well as to the adaption of the original Transformer model [2], to achieve performance similar to that reported by the authors. From our perspective, we cannot tell whether or not better PYTHON programming skills would have been beneficial and if someone with more experience with the multi-head attention would have been able to understand and implement the method right away. In any case, the authors certainly developed a coherent methodology and, by providing the

[Re] Satellite Image Time Series Classification with Pixel-Set Encoders and Temporal Self-Attention

Table 2. Results of the proposed method and our adapted version of it when being trained on SI-T33TWM. Generally, we evaluated four different scenarios where we used two different label files with each time two different ways of splitting the test set from the training set. The labels either represented the top 20 crop type classes found in the FR-T31TFM dataset (top-20-F) or the top 20 classes from the region of SI-T33TWM (top-20-S). As we set the parcels from one region of Slovenia aside, we were able to evaluate the methods not only on the proposed randomly drawn, but also on a regionally separated test set. It is necessary to recollect that with the stated top-20-F classes the entire dataset has less crop types to classify, which is illustrated in ??.

| Model variant | top-20-F classes | | top-20-S classes | |
|------------------------------|------------------|----------------|------------------|----------------|
| | random split | regional split | random split | regional split |
| TAE [1] | 90.92 | 89.80 | 87.42 | 83.84 |
| TAE with FC _v [2] | 90.88 | 89.50 | 87.50 | 83.86 |

corresponding code alongside with the paper, have ensured that all interested parties can clearly follow their ideas.

As a result of our misinterpretation of the Transformer adaptation proposed by Garnot *et al.*, we came across an alternative approach that performed comparably well as the one from the paper, but requiring only 80 % of its parameters. When we asked the authors about this observation, we concluded that we all had different opinions on the most obvious derivative of the method developed by Vaswani *et al.* [2] applied to the problem of satellite time series classification. Upon request, the authors of the Transformer paper confirmed that keeping the fully-connected layer FC_v has proven to be helpful and acknowledged the validity of our approach. Under these circumstances, it is not straightforwardly answerable whether the implicit claim of having the Transformer adapted in the papers' way can be supported.

However, it can be said that the well-documented code and clean GITHUB repository contributed strongly to our understanding of the method and helped us answer most of our comprehension question. Since this all is publicly available, a reproduction of the presented method based on that implementation is possible.

3.2 Reproducibility of the outcomes

Besides the possibility to reproduce the methodological aspects of the original paper, we were also interested in whether we could achieve the results stated in the investigated paper. By training the entire model by ourselves with the original dataset, we faced no considerable difficulties using the data. We were, in fact, able to achieve a slightly higher overall accuracy on the original test set compared to using the model pre-trained by the authors.

However, a slight drop in performance became observable when we used an alternative yet similarly preprocessed dataset. While we still reached an overall accuracy of over 87 % for the random split on our new and potentially more challenging dataset, we were not able to reach 84 % when running the inference on a regionally separated test set. This result highlights the importance of the right choice of a representative test set, especially when using Earth observation imagery, while still acknowledging that the presented method is potentially able to generalise to unseen and new data. Nevertheless, our reproduction experiments confirmed the claimed validity of the approach and it is to be left to the authors and the entire research community to investigate whether the presented model is able to compete with state-of-the-art methods given broader and more diverse datasets.

The only issue with the presented method arose when we used one dataset for training and another one for testing. Although we tried to preprocess our new dataset exactly the way the authors did, we were not able to obtain convincing results. There might be

several reasons causing this issue, like an incorrect harmonisation of the dataset, other parameters for the interpolation of the cloud-covered pixels, or assumptions about the data we unknowingly and implicitly made different than the authors. Hence, it remains interesting for us to know how exactly the data was processed.

To summarise, we support the claim that the method can successfully classify crop parcels when it was trained on data that was acquired under the same conditions, as the data it eventually gets tested on.

4 Conclusion

During the proposed study of “Satellite Image Time Series Classification with Pixel-Set Encoders and Temporal Self-Attention” by Garnot *et al.* [1], we assessed several questions regarding different aspects of the reproducibility of the paper. Therefore, we first attempted to reimplement the methodology from scratch based on the descriptions given in that paper. As this proved to be more challenging than expected and prone to misunderstandings, we proceeded to evaluate the provided clean implementation in terms of an adaption of the Transformer architecture [2]. There, we came across a discrepancy between our understanding of the vanilla multi-head attention concept and the one used in the paper. Our obtained results show that, by changing the proposed adaptation in a subtle way towards the more basic multi-head attention, the model uses considerably fewer parameters, while still performing equally well.

When employing the authors’ implementation and dataset, we were able to reproduce the presented results straightforwardly and even on a new dataset that we specifically developed for this survey, the approach delivered meaningful results. The only issues arose when the training and the test dataset did not share exactly the same properties lifting the accurate preprocessing of the data to a crucial component of the proposed method.

In conclusion, we can state that the examined method is conclusive in itself and valid. Our experiments speak in favour of the approach and our findings might highlight a path in which further works should proceed. This direction of research could take advantage of the dataset which we will make publicly available within the next few months.

References

1. V. S. F. Garnot, L. Landrieu, S. Giordano, and N. Chehata. “Satellite Image Time Series Classification with Pixel-Set Encoders and Temporal Self-Attention.” In: **IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)**. 2020, pp. 12322–12331. doi: 10.1109/CVPR42600.2020.01234.
2. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. “Attention is All you Need.” In: **Advances in Neural Information Processing Systems (NIPS)**. Vol. 30. 2017, pp. 5998–6008.
3. D. J. Lary *et al.* “Machine Learning Applications for Earth Observation.” In: **Earth Observation Open Science and Innovation**. Ed. by P.-P. Mathieu and C. Aubrecht. Vol. 15. ISSI Scientific Report Series. Springer International Publishing, 2018, pp. 165–218. doi: 10.1007/978-3-319-65633-5_8.
4. A. E. Maxwell, T. A. Warner, and F. Fang. “Implementation of machine-learning classification in remote sensing: an applied review.” In: **International Journal of Remote Sensing** 39.9 (2018), pp. 2784–2817. doi: 10.1080/01431161.2018.1433343.
5. J. S. Spiker and T. A. Warner. “Scale and Spatial Autocorrelation From A Remote Sensing Perspective.” In: **Geo-Spatial Technologies in Urban Environments: Policy, Practice, and Pixels**. Ed. by R. R. Jensen, J. D. Gatrell, and D. McLean. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007. Chap. 10, pp. 197–213. doi: 10.1007/978-3-540-69417-5_10.
6. A. D’Amour *et al.* “Underspecification Presents Challenges for Credibility in Modern Machine Learning.” In: (Nov. 6, 2020). arXiv:2011.03395 [cs.LG, stat.ML].
7. A. Paszke *et al.* “PyTorch: An Imperative Style, High-Performance Deep Learning Library.” In: **Advances in Neural Information Processing Systems (NeurIPS)**. Vol. 32. 2019, pp. 8024–8035.

4 Challenges and Opportunities of Large Transnational Datasets: A Case Study on European Administrative Crop Data

Maja Schneider, Christian Marchington and Marco Körner (2022). 'Challenges and Opportunities of Large Transnational Datasets: A Case Study on European Administrative Crop Data'. In: *Workshop on Broadening Research Collaborations in ML (NeurIPS 2022)*. DOI: [10.48550/arXiv.2210.07178](https://doi.org/10.48550/arXiv.2210.07178)

Challenges and Opportunities of Large Transnational Datasets: A Case Study on European Administrative Crop Data

Maja Schneider

Technical University of Munich (TUM)
TUM School of Engineering and Design
80333 Munich, Germany
maja.schneider@tum.de

Christian Marchington

London, UK
christian@marchington.dev

Marco Körner

Technical University of Munich (TUM)
TUM School of Engineering and Design
80333 Munich, Germany
marco.koerner@tum.de

Abstract

Expansive, informative datasets are vital in providing foundations and possibilities for scientific research and development across many fields of study. Assembly of grand datasets, however, frequently poses difficulty for the author and stakeholders alike, with a variety of considerations required throughout the collaboration efforts and development lifecycle. In this work, we discuss and analyse the challenges and opportunities we faced throughout the creation of a transnational, European agricultural dataset containing reference labels of cultivated crops. Together, this forms a succinct framework of important elements one should consider when forging a dataset of their own.

1 Introduction

With the progression of member states of the *European Union (EU)* to continually and openly publish administrative data, the opportunity to conduct research previously limited to a local-scale has advanced massively with the emergence of expansive, collaborative, and multi-national datasets. The EuroCrops initiative [6, 7, 9] harnessed the continually improving accessibility to *Common agriculture policy (CAP)* [3] data, demonstrating the feasibility of a transnational dataset and building out a framework of considerations one must account for when developing projects at a similar spatial coverage. While there are clear advantages of a large and diverse dataset for all breadths of research, the number of challenges imposed by administrations' legacy, and often manual, systems continues to make gathering data at this scale an ever-laborious task. Marini and Nicolardi [4], for instance, face these obstacles when they harmonised several sub-databases containing real estate information in order to analyse the social and economic changes within Europe. More generally, Connelly et al. [1] analysed the issues with administrative social science data while stressing the importance of its use.

In this paper, we distilled the challenges we faced while working on EuroCrops into a framework consisting of six distinct categories, which will be further explored in the sections following. The goal of this framework is to provide researchers and authorities with guidelines when approaching transnational, country-dependent, and collaborative open data projects.

Workshop on Broadening Research Collaborations in ML (NeurIPS 2022).

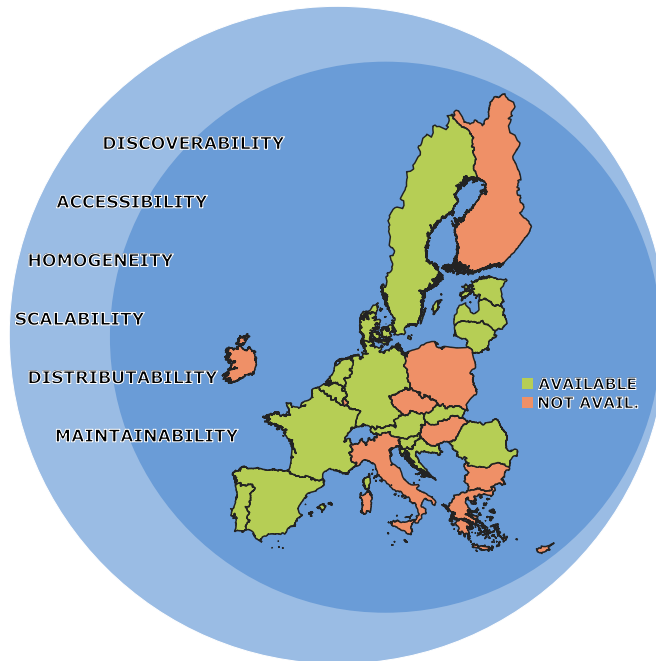


Figure 1: The EuroCrops dataset now includes data from 16 member states of the *European Union (EU)*. The countries coloured in green provide data and helped with the identification of the six biggest challenges faced when compiling collaborative, transnational datasets, which are listed on the left. At the time of the first EuroCrops release, data from the red countries was not available yet.

While highlighting the challenges involved, we also want to encourage the community to take part in the development of large pan-European datasets by showcasing the advantages and benefits they are able to provide for a multitude of stakeholders.

2 The Study Data

EuroCrops can be considered to be the first large-scale, pan-European dataset for cultivated crops, following the approaches of the LUCAS [2] and BreizhCrops [5] initiatives, combining and harmonising georeferenced agricultural parcel data with the corresponding crop. While originally motivated as a use-case for applying machine learning methods to Earth observation data, more specifically Copernicus Sentinel-2 satellite imagery, EuroCrops has now become an impactful project spanning across a number of disciplines, including biodiversity and agriculture. The data itself was obtained through the countries' agricultural ministries or paying agencies and contains the farmers' self-declarations, which are collected within the subsidy control of the *Common agriculture policy (CAP)*. At the time of this study's publication, 16 member states of the *EU* have contributed data, highlighted in green in Figure 1, which we collected over a span of 1.5 years.

As part of the collection effort, a new *Hierarchical Crop and Agriculture Taxonomy (HCAT)* [8] was designed to both accommodate the high crop diversity and harmonise country-unique cultivation schemas and languages across Europe. The harmonised EuroCrops vector data is centrally and

publicly available via the project's GitHub repository¹, on Zenodo², and will be distributed over commonly utilised platforms, such as Google Earth Engine³, GeoDB⁴, CODE-DE⁵, and EO-Lab⁶.

3 Challenges

While compiling the EuroCrops dataset, we identified the following six challenges, forming the foundation of the aforementioned framework.

3.1 Discoverability

One would expect that "open" data would indicate the simplicity of locating the hosted data and determining the type and quantity of the available data; we found this was often not the case. As countries frequently choose not to use central data distribution points but instead their own national platforms, the method of actually finding these self-hosted repositories still requires incredibly manual investigatory work. Similar data is often hosted by different types of responsible agencies and authorities across Europe, raising the barrier when attempting to discover the available sources.

Even in the scenarios where the correct authority contact and platform is determined, one still encounters language barriers when navigating local websites, as they are often only hosted in their native language.

3.2 Accessibility

The ever-growing concerns surrounding security and storage of personal data continues to motivate authorities and platforms to restrict direct access to the data itself. This is especially true for data that can be linked back to individuals, notably farmers in the case of EuroCrops, and provides authorities the justification to obscure themselves behind GDPR walls, even if there are legally-just reasons for public access.

Unfortunately, the benefits of taking the step to openly release the data are not justifiably strong enough for the responsible parties to undertake such an effort. From the outside, there is no observably clear argument against working on the publication, and we can only hypothesise that it might be connected to lack of digitalisation efforts, internal politics, and availability of resources to appropriately prepare the data. Where resource has been found and appropriately applied, it has sometimes resulted in platforms necessitating user registration before being able to gain access, further adding to the manual effort.

3.3 Homogeneity

While lightly touched on previously, an important consideration is the variability in the data content released across countries. Regarding cultivated vegetation, each country groups their crop data into unique taxonomies and schemes, which further vary between legacy and modern formats. This includes mixtures of scientific plant nomenclature and terminology in national languages, but also the depth of granularity, ranging from precise biological species to the more generic plant families.

Not only does the data content vary across countries within the EU, but so do the data formats, most notably in fields where severally accepted formats compete within the community. For example, we observed that, despite being proprietary, Esri shape files still contribute to a large proportion of the geospatial data pool, amongst more modern formats, such as geopackage or webmapservice. The necessity to handle multiple, conflicting formats when pulling the data from respective authorities further impedes research and development at a transnational scale.

¹<https://github.com/maja601/EuroCrops>.

²<https://doi.org/10.5281/zenodo.6866846>.

³<https://earthengine.google.com/>.

⁴https://eurodatacube.com/marketplace/services/edc_geodb.

⁵<https://code-de.org/en/>.

⁶<https://eo-lab.org/en/>.

3.4 Scalability

In the cases where individual country-dependent datasets are fetched, harmonised, and combined into an aggregated collection, the collection itself is still encapsulated within the pool of available national datasets. The effort of including new sources then requires restarting this harmonisation process from the beginning, to ensure the content of the additional data complies with the existing collection, while the existing structures have to be altered to accommodate the new dataset. This is not only true when new countries are added to the transnational collection, but also when ministries or authorities of a certain country migrate from legacy data platforms, or update content structure or format. These can happen irregularly over the course of years and therefore require manual monitoring and intervention.

3.5 Distributability

When releasing or updating the complete dataset, there is the problem of ensuring that all relevant stakeholders receive notification and central access to the data. Under current circumstances, this typically involves reaching out to each member state individually. Even though European data repositories exist, these currently lack the functionality to act as a distribution hub, where all information is brought together and subsequently used as an exchange platform for all concerned parties: it is still required to actively engage authorities to ensure they receive the latest updates.

3.6 Maintainability

With the increasing size and number of interactions with a dataset, progressively entangled, moving parts begin to co-exist and need to be continuously maintained. In the case of EuroCrops, a number of country-owned variables will undergo constant development, including: changes to crop policies; infrastructure updates, such as to platforms and authority websites; and standards. Each of these provides some underlying structure or metadata to the dataset itself, and so require strict manual verification to ensure a high dataset quality is maintained.

When changes to country-owned variables do occur, these are typically carried out without communication, version-control, and tend to be obscured in the background, with error corrections even understated on the authorities' data platforms. This makes it extremely challenging to pin down and correct outliers in datasets without significant effort.

4 Opportunities

While construction of these large, interwoven datasets presents the array of challenges as discussed above, the opportunities of providing communities with dataset like EuroCrops far outweigh the effort involved. In the following section, we discuss three of the key target groups and the complementary benefits that the dataset would deliver.

4.1 For Earth Observation

As many in the Earth observation community are aware, provision to easily-accessible and complete datasets is often limited and difficult to obtain. Initiatives to create open-access datasets covering vast swathes of Europe considerably lowers the barrier to entry, providing members of the community with large-scale, manipulable datasets on which they can begin to operate. These datasets are important, as large-scale Earth observation analysis might require substantially more reference ground data for training machine learning models than one country may be able to provide alone.

Furthermore, Earth observation reference data is often complicated to both understand and work with, being frequently obscured by multiple layers of abstraction and more difficult to find when compared to natural image labels. By overcoming this hurdle, the domain itself sees further benefits through increased accessibility to outside researchers.

4.2 For Europe

A clear benefit of transnational datasets is the weakening of countries' data sovereignty, which has historically led to encapsulated and incompatible granular datasets and models.

By giving the citizens of the European Union the ability to access and use all available data in a consolidated location and format, research can take a transnational direction, allowing for broader, pan-European problems to be tackled and solutions developed. This in turn enables far greater potential for both bi- and multilateral research projects and data analysis across borders.

4.3 For Ministries and Authorities

When considering ministries and authorities, the benefits of actively participating within transnational datasets take another clear step: data can be accessed, controlled, and maintained within a singular ecosystem; changes can be version-controlled and publicly-stated; and metadata can be standardised.

By providing a unified node that holds all necessary information to download and work with the dataset, authorities are provided with a simple means of interacting with the public and those who request access, as well as establishing a forum for discussion and feedback to be held.

Additionally, with the growth of a singular, established, transnational dataset, there comes the appeal for other countries to participate. This provides consequent and progressive reward to participants, as they can rely on an existing baseline against which they can fit their data, instead of developing another from the ground up. Making the entire process far more scalable, both horizontally and vertically, delivers tangible time and effort benefits from the start.

5 Conclusion

In this article, we have presented and discussed both the challenges and opportunities that the development of a large-scale, transnational dataset may deliver.

Through the introduction of EuroCrops, we have started to tackle the *discoverability* and *accessibility* barriers by providing one publicly available dataset, which is currently being published on several, well-established and easily discoverable platforms. The development of *HCAT* expands on this, introducing a sense of *homogeneity* through harmonisation of reference data and allowing for *scalability*, both with new data from existing countries, as well as newly participating countries.

Finally, during the project itself, we managed to build up a community of data providers and data users and form stable connections with a number of authorities and paying agencies across several member states of the European Union. This way, we hope to actively address the issue of *maintainability* of the dataset and *distributability* of information to all participants of the process, through our now established lines of communication and continually addressing feedback and concerns.

References

- [1] R. Connelly, C. J. Playford, V. Gayle, and C. Dibben. The role of administrative data in the big data revolution in social science research. 59:1–12. doi: 10.1016/j.ssresearch.2016.04.015.
- [2] R. d’Andrimont, M. Yordanov, L. Martinez-Sanchez, B. Eiselt, A. Palmieri, P. Dominici, J. Gallego, H. I. Reuter, C. Joebges, G. Lemoine, et al. Harmonised lucas in-situ land cover and use database for field surveys from 2006 to 2018 in the european union. *Scientific data*, 7(1):1–15, 2020.
- [3] European Commission. Common agricultural policy. *agriculture.ec.europa.eu* https://agriculture.ec.europa.eu/common-agricultural-policy_en.
- [4] C. Marini and V. Nicolardi. Big data and economic analysis: The challenge of a harmonized database. In *Studies in Classification, Data Analysis, and Knowledge Organization*, pages 235–246. Springer. doi: 10.1007/978-3-030-51222-4_18.
- [5] M. Rußwurm, C. Pelletier, M. Zollner, S. Lefèvre, and M. Körner. Breizhcrops: A time series dataset for crop type mapping. *arXiv preprint arXiv:1905.11893*, 2019.
- [6] M. Schneider and M. Körner. TinyEuroCrops. *mediaTUM* doi.org/10.14459/2021MP1615987, 2021. Dataset.
- [7] M. Schneider and M. Körner. EuroCrops. *Zenodo* doi.org/10.5281/zenodo.6866846, 2022. Dataset.

- [8] M. Schneider, D. Gackstetter, J. Prexl, S. T. Meyer, and M. Körner. Analysing the Impact of European Agriculture on Biodiversity with an updated Hierarchical Crop and Agriculture Taxonomy (HCAV2). Manuscript in preparation.
- [9] M. Schneider, A. Broszeit, and M. Körner. EuroCrops: A pan-european dataset for time series crop type classification. In P. Soille, S. Loekken, and S. Albani, editors, *Proceedings of the Conference on Big Data from Space (BiDS)*. Publications Office of the European Union, 2021. doi: 10.2760/125905.

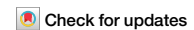
5 Advancing Transnational Assessments of Biodiversity Drivers in European Agriculture with an Updated Hierarchical Crop and Agriculture Taxonomy (HCAT)

Maja Schneider, David Gackstetter, Jonathan Prexl, Sebastian T. Meyer and Marco Körner (2025). 'Advancing Transnational Assessments of Biodiversity Drivers in European Agriculture with an Updated Hierarchical Crop and Agriculture Taxonomy (HCAT)'. in: *npj Sustainable Agriculture* 3.1, pp. 1–10. DOI: [10.1038/s44264-024-00037-x](https://doi.org/10.1038/s44264-024-00037-x)



<https://doi.org/10.1038/s44264-024-00037-x>

Advancing transnational assessments of biodiversity drivers in European agriculture with an updated hierarchical crop and agriculture taxonomy (HCAT)



Maja Schneider^{1,6}, David Gackstetter^{1,2,3,6}✉, Jonathan Prexl⁴, Sebastian T. Meyer⁵ & Marco Körner¹

By homogenizing landscapes and reducing natural habitats, modern agriculture plays a significant role in reducing natural species populations worldwide. Despite advances in research, quantifying the impacts of cropping systems on biodiversity remains challenging due to the lack of comprehensive agricultural data. Within the European Union's (EU) common agricultural policy (CAP), farmers are required to declare cropping arrangements to receive subsidies. The resulting data is collected by each EU member state individually, leading to inconsistent crop taxonomies across the EU, which hinders transnational analyses of agriculture and related impacts. To overcome this barrier, we developed a hierarchical crop and agriculture taxonomy (HCAT), which harmonizes administrative and agricultural data from 16 EU member states. With the release of an upgraded second version of HCAT, we demonstrate, using the example of biodiversity drivers, how a harmonized CAP data set can aid in identifying indicators related to environmental impacts in agricultural landscapes at international scales.

Since the 1960s, global agricultural production has increased by 2–2.5% per year. The main drivers for this development refer to additional land use, input intensification, breeding advances, and efficiency gains from technological innovations. At the same time, undesirable externalities of farming practices, such as water pollution, soil degradation, and greenhouse gas emissions, have drastically increased^{1,2}. For decades now, unprecedented rates of biodiversity loss have been recorded globally^{3,4}. Areas with more intense agriculture practices have thereby been associated with greater losses in biodiversity⁵.

One important sub-aspect characterizing the intensity refers to the shaping and arranging of the agricultural and natural landscape for productive farming. Historic developments such as the spatial separation of arable crops from animal husbandry, the diminishing diversity of cultivated crops, and the conversion of semi-natural land into crop fields have changed agricultural landscapes, particularly since the beginning of the 20th century^{6,7}. Next to changes in landscape composition, also the spatial

arrangement and configuration of agricultural landscapes altered. Factors like labor rationalization through mechanization or land consolidation programs^{8,9} led to increases in field sizes and reductions of field margins (e.g., grass trips and hedgerows), which caused the “mosaics of fields”¹⁰ to become more coarse-grained and monotonous with less diverse resources. Together with extensive transformations of natural areas into agricultural land, the clearance of the landscape and its entailed division into natural and crop management-related areas severely reduced the connectivity of ecosystems. This represents another driver for the degradation of functional ecosystems, as natural species—especially mobile species—significantly benefit from the interconnection of ecosystems^{6,10–12}.

Beyond pure spatial characteristics of agricultural landscapes, knowledge about the types of individual crops is another important information for ecological assessments. Already at the single-field scale, crop types can have varying impacts on the surrounding natural ecosystems. When grown in large areas, oilseed rape, for instance, is shown to weaken pollinator

¹Chair of Remote Sensing Technology, TUM School of Engineering and Design, Technical University of Munich, 80333 Munich, Germany. ²Hans Eisenmann-Forum for Agricultural Sciences (TUM-HEF), Technical University of Munich, 80333 Munich, Germany. ³Precision Agriculture Lab, Department for Life Science Engineering, TUM School of Life Sciences, Technical University of Munich, 80333 Munich, Germany. ⁴Department of Aerospace Engineering, University of the Bundeswehr Munich, 85579 Neubiberg, Germany. ⁵Terrestrial Ecology Research Group, Department for Life Science Systems, TUM School of Life Sciences, Technical University of Munich, 80333 Munich, Germany. ⁶These authors contributed equally: Maja Schneider, David Gackstetter. ✉e-mail: david.gackstetter@tum.de

richness independently of the semi-natural areas in the surrounding landscape¹³. In contrast, late-flowering crops (e.g., clover) may provide valuable resources for wild pollinators late in the season¹⁴. Beyond assessments on the single-field level, recent years have produced significant progress on the landscape scale investigating the effects of varying crop types across several fields and surrounding semi-natural land. In 2014, Palmu et al. investigated the relationship between landscape-scale crop diversity, farming practices, and ground beetle diversity in Southern Sweden. They discovered a positive correlation between crop diversity and ground beetle richness, which was influenced by different levels of agricultural management intensity¹⁵. Aguilera et al.¹⁶ confirmed the positive correlation between crop diversity and natural biodiversity, additionally highlighting the positive influence of abundant semi-natural habitats¹⁶. Other studies focused on the geometry and arrangement of the cropland and semi-natural areas within a landscape. These showed the positive effects of more heterogenic and small-scale arrangements of cropland on non-crop abundance and diversity – even without creating additional field borders, or semi-natural habitats, i.e., without taking land out of agricultural production^{6,17,18}. The latter finding is particularly remarkable when it comes to balancing the conflict of goals between productivity and sustainability.

These studies highlight that having detailed information on the presence, diversity, and spatiotemporal arrangement of cultivated crops in a landscape can be a precious information source for explaining conditions and variations of ecosystems. However, when not collecting field data for country-overarching studies but using available data sources^{17,19}, study areas are often limited by regional or national political borders. Consequently, today's literature on transnational, cross-border ecosystems affected by cropland is still limited despite the necessity for a better, comprehensive understanding of the complex interrelations between crop management practices and biological diversity^{20–23}. This knowledge would be fundamental for developing more holistic, transnationally coordinated policy recommendations on balancing crop production and environmental protection²⁰. The two primary reasons for the limited study areas in this context refer hereby to (1) the varying data access policy and (2) the lack of standardization of crop class nomenclature between countries³⁵. Researchers have tried to bypass these limitations by either focusing on only one country at a time^{25,26} or by significantly reducing the number of classes²⁵.

In the European Union (EU), farmers are required to annually declare every crop that they grow on each field to receive subsidies within the framework of the common agricultural policy (CAP)^{27,28}. This data is centrally collected in the EU's Integrated Administration and Control System (IACS)²⁹. This procedure produces data covering millions of geo-referenced fields across the entire EU, providing information on the exact location of fields and their respective types of cultivated crops. Even though the data is collected on the EU level, the data access policy is defined by each member country individually. Over the past years, several countries have decided to make their data publicly available, resulting in a leap of data-driven crop type monitoring with modern machine learning techniques and research in large-scale vegetation analysis^{25,30,31}. Unfortunately, not all member states have decided to take this opening step, and even if the data is published, each country utilizes different national crop schemas, taxonomies, and formats. Hence, transnational aggregations and comparisons become impossible if the data is not pre-processed and harmonized beforehand. Limiting study areas to only individual countries or a few manually selected, laboriously pre-processed crop sub-collections strongly hinders the research community from exploiting the full potential of the EU's multinational crop data set. Transnational crop taxonomies are thereby identified as suitable means to overcome the abovementioned barriers. There are indeed schemes for the classification of land cover and land use on a European level, such as EAGLE or CORINE^{32,33}, which, however, lack detail when it comes to fine distinctions of agricultural crops. The Indicative Crop Classification (ICC) taxonomy of the United Nations Food and Agriculture Organization (FAO) depicts a general crop classification scheme, which hierarchically structures a wide range of agricultural crops and specifies related characteristics, including specific crop genus or species, product type, and growing cycle

(temporary/permanent)³⁴. However, despite the apparent necessity for multinational analysis and research, there is, to our knowledge today, still no structured and consistent transnational taxonomy specifically for administrative crop declarations that enable the translation of national-specific crop types into transnationally harmonized crop notations.

The EuroCrops project represents an important initiative in this context, as it aims to build an EU-wide transnational data set for IACS data^{34,35}. One key element of this project represents the development of a harmonized crop taxonomy to bridge the abovementioned barriers and to advance the transnationally seamless usage of the data. In 2021 the project presented the first prototypical Hierarchical Crop and Agriculture Taxonomy version 1 (HCATv1)³⁴. As this taxonomy was evaluated as not universal enough for cross-discipline purposes, we developed an updated transnational crop taxonomy, denoted Hierarchical Crop and Agriculture Taxonomy version 2 (HCATv2 or just HCAT). This renewed version aims to tackle the abovementioned challenges by harmonizing all crop classes in the datasets obtained from the authorities and paying agencies across the EU. HCATv2 enables cross-national research in agriculture and related fields while letting individuals choose the desired scale. Specifically, for agroecological research, HCAT-based datasets can now facilitate transnational comparisons and studies of cross-border ecosystems influenced by crop management activities in the EU. The aim of this work is (1) to discuss the common European crop taxonomy HCAT including the harmonization of administrative, and agricultural data from 16 EU member states within the already existing common dataset collection of publicly available national crop declarations being collected in the framework of EuroCrops³⁵, and (2) to investigate the potentials of HCAT for different applications domains, with a special focus on ecological research in the context of agricultural landscapes, among others by demonstrating an exemplary case study.

Results

Hierarchical crop and agriculture taxonomy (HCAT)

The hierarchical crop and agriculture taxonomy version 2 (HCATv2 or HCAT) is a tree scheme structuring all crop classes present in publicly available European IACS (or Land Parcel Identification System (LPIS)) datasets into six levels (Fig. 1). The classification approach generally follows the biological grouping of crop species. Each HCAT class consists of a name in the English language and an identifier; the latter indicates the level of the current crop in the hierarchy. This way, it is possible to classify, e.g., winter barley (33-01-01-04-01) as a subclass of barley (33-01-01-04-00), which itself is a subclass of cereal (33-01-01-00-00). Level 2 (root node in Figs. 1) and 1 are given by the position of HCAT within the EAGLE classification scheme for land use and land cover³² and are kept constant with prefix 33 in HCAT. More details on the development process of HCAT can be retrieved from the Method section as well as from previous publications on HCAT and EuroCrops^{34,35}.

Descriptive statistics of HCAT

Each sub-dataset in the EuroCrops dataset collection of EU countries' crop declarations comes with an inherent level of detail of crop type denominations³⁵. Within the scope of data from 16 EU member states (Fig. 2), the maximum numbers of available crop classes range from 15 in Croatia to 326 in the Netherlands (Supplementary Table 1). As this is the original data from the countries, that is the finest possible distinction between crops. Harmonizing the respective datasets entailed a loss of information as the definition of classes does not address each country's minor, very specific differentiations of crops (e.g., types of meadow). Building upon HCATv2 taxonomy and the scope of data providing countries (Supplementary Table 2), Supplementary Tables 3–5 show the bilateral comparisons between each of the participating countries for the three coarsest levels of the taxonomy, i.e., level 6 to level 4. Each number in the grid represents the

<https://doi.org/10.1038/s44264-024-00037-x>

Article



Fig. 1 | Multi-layer hierarchy of HCATv2. Description: All crop classes present in HCAT adhere to an internal, hierarchical structure. This hierarchy is shown here by a radial dendrogram where the start is in the center, and each layer represents one level of the taxonomy. The outer circle shows level 6 with the highest degree of detail to the type of crop. Levels 5 and 4 are represented by the circular representations

going inwards. Level 3 is the coarsest level that still differentiates between crop type classes. Levels 2 and 1 are fixed across all HCAT classes and given by the position of crop types within the EAGLE land use and land cover taxonomy³² and summarized here visually in the root node (“HCAT”). This dendrogram’s contents, including complete crop-type denominations, can be investigated in Supplementary Table 1.

count of mutually common crop classes of two countries, except the diagonals, which show the total number of crop classes of one country. With decreasing HCAT level, i.e., by reducing the level of detail in crop type denominations, the numbers of crop classes per country as well as the absolute frequency of mutually common crop classes between two countries decreases. Simultaneously, the relative frequency of common HCAT crop classes increases between

countries and, by that, the degree of comparability. We define comparability hereby as the ratio of countries’ mutually present HCAT classes to the total number of classes of the considered countries. This trend is visualized in Fig. 3, where yellow to white connections indicate high similarities of crop taxonomies between two countries, while red connections indicate low similarities. With an increasing share of common HCAT classes between countries, the

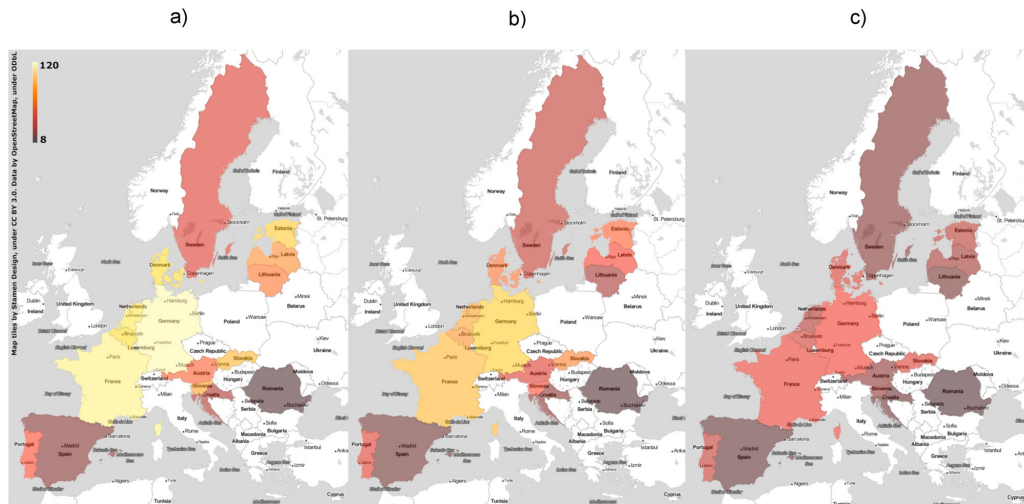


Fig. 2 | Number of differentiable crop classes per HCAT level across the considered scope of EU countries and regions. Description: These maps show each state's number of differentiable crop classes when considering HCAT levels 6 (a), 5 (b), and 4 (c) (i.e., number on the diagonals of Supplementary Table 3, Supplementary Table 4, and Supplementary Table 5). A different way to obtain these numbers is by counting the maximum number of leaves (or most outer nodes) in the tree structure of Fig. 1 for each country. Level 6 includes all nodes; in case of level 5 and 4 the first respectively first and second outer rings are pruned off, and the resulting nodes represent the corresponding leaves. In some countries (e.g.,

Romania), most outer ring/level corresponds to level 4, which is why their counts don't change when adding or removing hierarchy levels 5 and 6. In the case of Belgium, Germany, and Spain, the data was taken from the subregions for which, at the time of development, data was already provided. In this visualization, the range of the legends stays the same across all three maps, highlighting the increasing similarity of the set of crop types with decreasing HCAT levels. If the highest level of detail is chosen, the heterogeneity in sets of crop types is at its maximum, as shown in map (a). The more homogeneous crop taxonomies at level 4 (map (c)), enable larger spatial coverage in transnational analysis.

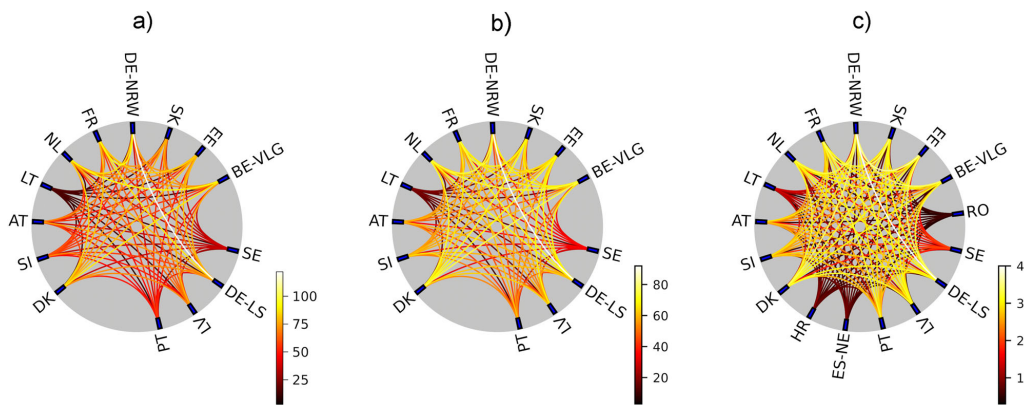


Fig. 3 | Relative share of common crop classes per HCAT level across the considered scope of EU countries and regions. Description: The listed numbers of similar classes between two countries for HCAT levels 6 (a), 5 (b), and 4 (c) (corresponding to Supplementary Table 3, Supplementary Table 4, and Supplementary Table 5) are visualized in three-chord diagrams: Each bilateral connection is colored according to the similarity of crop taxonomy between the respective countries (EE

Estland, SK Slovakia, DE-NRW North Rhine-Westfalia in Germany, FR France, NL Netherlands, LT Lithuania, AT Austria, SI Slovenia, DK Denmark, HR Croatia, ES-NA Navarra in Spain, PT Portugal, LV Latvia, DE-LS Lower Saxony in Germany, SE Sweden, RO Rumania, BE-VLG Flanders in Belgium). Yellow-to-white connections indicate high shares of common crop types in two countries' taxonomies; red-to-black connections imply lower degrees of similarity.

number of field parcels, respectively, and the size of the area that can be considered for analysis increases.

Case study

To showcase the potential of HCAT in facilitating research on crop-related diversity, we conducted a case study comparing crop diversity

between exemplary regions at different HCAT levels (see Methods). For seven regions and the HCAT levels 4, 5, and 6, we calculated Shannon diversity H_i and the number of crop types $n_{c,i}$ for grid cells of 1 km × 1 km (Supplementary Fig. 1 and Supplementary Fig. 2). Having applied the same color scale across each metric, it can be observed that on average both the Shannon index H_i and the number of crop types $n_{c,i}$ decrease

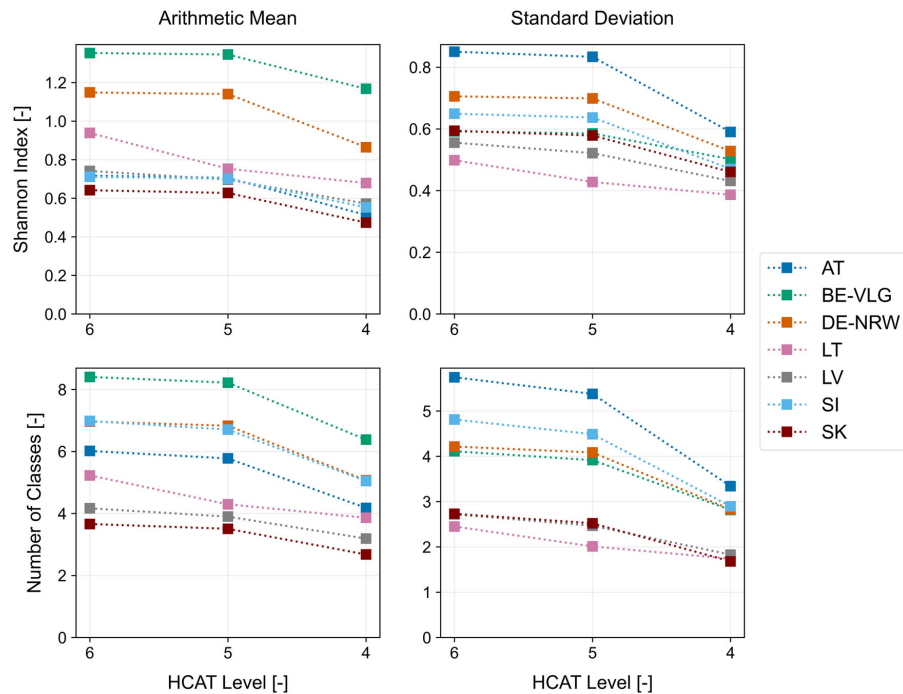


Fig. 4 | Comparing trends in crop diversity of selected EU regions and countries based on Shannon index H_i and of number of classes $n_{c,i}$. Description: Line graphs visualizing crop diversity per unit of area using arithmetic mean and standard deviation of the Shannon index H_i and of number of classes $n_{c,i}$ per country and

HCAT-level (4, 5, and 6) for seven considered regions/countries (AT Austria, DE-NRW North Rhine-Westphalia in Germany, LT Lithuania, LV Latvia, SI Slovenia, SK Slovakia, BE-VLG Flanders in Belgium).

with a lowering HCAT level. Further, we can see large differences for both metrics within (e.g., Austria and between regions (e.g., Navarra compared to Slovenia). The magnitude of these intra- and transregional differences appears to reduce, i.e., the regions increasingly assimilate and homogenize, with decreasing HCAT levels. These visual trends can also be quantitatively observed for H_i and $n_{c,i}$ in Supplementary Fig. 3 and Supplementary Fig. 4. With a few exceptions, we can see similar trends across the two considered metrics: Flanders stands out with the highest arithmetic mean value compared to Slovakia with the lowest values (Fig. 4, Supplementary Figs. 3 and 4). Austria shows the highest variability within the country, given in terms of the standard deviation and the extended interquartile range (whiskers range in boxplots). Figure 4 also quantitatively confirms the above-mentioned qualitative observation of a generally lowering heterogeneity between countries with decreasing HCAT levels. There's also a distinct exception observable from the overall similar trends for the two metrics: Austria and Slovenia stand out with significantly lower mean values for H_i than for $n_{c,i}$ compared to the other regions.

The number of crop types $n_{c,i}$ represents an important factor for the H_i index. Yet beyond, also the size and number of fields per area (here discrete grid cells) influence the H_i . Using Fig. 5 and Supplementary Fig. 5, we can observe that the regions differ partially significantly from each other with respect to the spatial characteristics of their agricultural landscapes, with some showing more than five times the average field sizes compared to other regions (e.g., Slovenia compared to Slovakia). The number of fields per grid cell generally shows an inverse trend compared to the median area of fields, i.e., regions with a relatively higher number of fields per grid cell show comparably smaller field sizes. It can further be observed that the number of crop types $n_{c,i}$ correlates with the number of fields per area with a mean

correlation coefficient of $r_{\text{mean}} = 0.70$, and a certain heterogeneity within the considered scope or regions ranging from $r_{\text{LV}} = 0.61$ to $r_{\text{SK}} = 0.76$.

Discussion

Despite the necessity of more quantitative information on farming activities and even though data on crop types are publicly available in the EU, there is today no structured taxonomy for administrative crop type notations across countries from multilingual language areas (see Introduction). With HCAT at hand, each EU country's original, georeferenced parcel data can be enriched towards an extended version that additionally includes a direct translation of the crop type and the corresponding assignment to its corresponding HCAT class with name and identifier. Such enriched, harmonized data sets overcome the persisting language barriers (due to national-specific crop declarations) and can then be combined for transnational analysis.

Due to its hierarchical structure, HCAT allows for high flexibility during study design. The finest hierarchical level keeps the highest degree of detail. Decreasing the HCAT level entails a reduction of detail in crop denominations, and consequently introduces a loss of information. At the same time, the share of similar classes between countries increases, which enlarges the spatial potential for transnational analysis. To date, some countries/regions (Croatia, Navarra (Spain), and Romania) have only reported their taxonomy at a degree of detail corresponding to HCAT level 4, which enables transnational analysis including these countries/regions only at this coarser level of detail. Depending on their research objectives, scientists must carefully consider in advance to which spatial extent they want to use the data and which specific crop types they target, and accordingly identify the suitable HCAT level. Having outlined the significant potentials of HCAT for transnational, ecological studies in the EU,

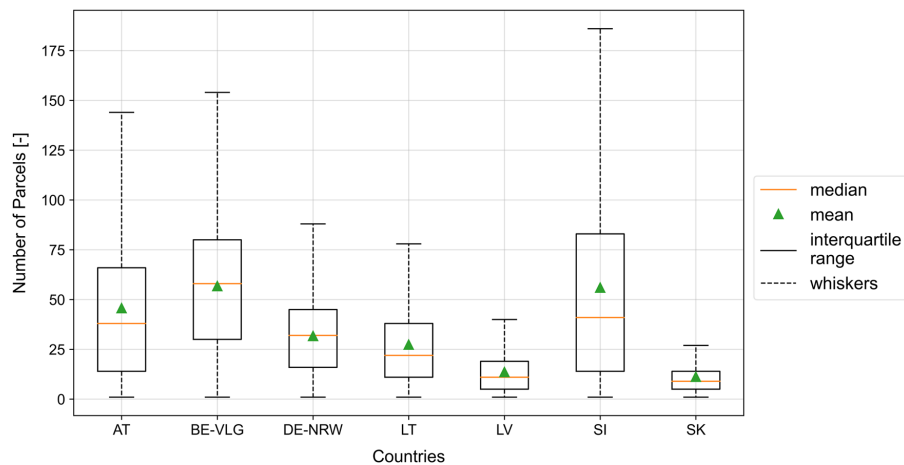


Fig. 5 | Comparing agricultural landscapes' spatial patterns of selected EU regions and countries based on the number of field parcels per defined area. Description: Boxplots showing mean (green triangle), median (orange line), 25–75% interquartile range (IQR) (black lined box), and whiskers value range (dashed lines),

defined as $IQR \pm 1.5$ times IQR, of a number of parcels for the defined unit of area for seven considered regions/countries (AT Austria, BE-VLG Flanders in Belgium, DE-NRW North Rhine-Westfalia in Germany, LT Lithuania, LV Latvia, SI Slovenia, SK Slovakia).

we, however, also want to discuss the currently given limitations and challenges of the taxonomy and the related usage of IACS data:

Despite IACS data often being utilized as reliable reference data for remote sensing applications^{25,30,31}, its limitations regarding correctness due to fraud or mistake need to be addressed and considered. Even though expert interviews with some control agencies in Germany have revealed empirical correctness levels above 95%, there is no official guarantee of EU-wide consistency in these terms. Consequently, analysis building upon the IACS data must be treated carefully, and methods applied that are robust to probably low but potentially existing declaration errors. Here the large size of the EU-wide EuroCrops data set offers an inherent countermeasure, as the relative importance of individual errors decreases with increasing data sets. Additionally, the subsidy does not rely on all crops cultivated over a year, so data collected within CAP usually only holds the primary or first crop grown on a particular parcel. Most of the time, this is enough information, but often, especially in warmer climates, it is known that there are several crop-growing cycles within one vegetation period. Any analysis that relies on the knowledge of conducted crop rotation is therefore limited to the presented type of data. Some member states like Austria and Portugal nevertheless have decided to release this information additionally, and even though we were not able to incorporate that into the vector data, we still attempted to provide translations to all specified crop types.

Regarding translations, it is also worth noting that even in the era of online dictionaries, most translation programs still struggle with national-specific agricultural terminology. This way, mistranslations soon became the central issue of the harmonization process and still hold the most significant potential of false classification of crops. Fortunately, several countries (Belgium, Croatia, Denmark, Estonia, France, Germany, Latvia, Portugal, and Slovenia) have decided to actively support us by providing corrections and feedback to our translation approach. Consequently, for these countries, the likeliness of mistranslations is at its lowest compared to other EU countries. The lowest error rates can also be expected for the most frequently and commonly cultivated crop types. To enable the possibility for individual quality assurance of HCAT-based experiments, we always provide the mapping of the HCAT classes next to the original, country-specific crop-type denominations.

Moreover, we also collected feedback from several collaborating institutions and potential users of HCAT regarding the taxonomy itself and provided them with discussion points for which we needed clarification. On

purpose, stakeholders were chosen from different disciplines to help build the bigger picture of requirements. As expected, the results were laying on the entire spectrum of answers which resulted in the acknowledgment that there was not this one singular way satisfying both, e.g., the remote sensing community uniting with their focus on similar hyperspectral reflectance values of crops and the biologists with their clear-defined families and species of plants. Eventually, we decided to go for a compromise that both worlds might not consider perfect but are anyways willing to use and integrate into their domain-specific schemes. Independent from the applied compromises, if required, user can always identify the initial crop types through the provided raw crop denomination in their original and English language.

Across the above-mentioned aspects, we can see that HCAT fundamentally depends on the collaboration of the individual member states regarding the public provision of data and clarifications of national, domain-specific knowledge. So far, the majority of collaborating states are located in central and northern Europe. Contributions from southern and eastern EU countries would, therefore, be particularly beneficial for advancing the spatial representativeness of HCAT across the entire EU in future versions. Besides the quantity of data, we also encourage each country/region to further revise and enhance their ontologies, especially those countries that still report only at a coarse level of detail (e.g., Croatia at HCAT level 4). Yet, it's also the public user and research community, which constantly contributes to improvements by providing feedback directly or via our community platform on GitHub³⁶. Among others, on this platform we provide latest updates on HCAT, users can share their suggestions and questions publicly via "Issues" or "Discussions", or directly propose edited file versions via Git. Being maintained still for several more years, the EuroCrops project warmly welcomes further reviews and contributions specifically for version 3, currently under development and generally for improved, future HCAT versions. By letting HCAT evolve over the past years and being open to discussions and feedback, it was possible to develop a scheme that shows the possibility of building a meaningful transnational taxonomy that holds more classes than all previous attempts. The hierarchical structure, open-data policy, and substantial spatial and linguistic coverage represent key characteristics of a scheme that is unprecedented in the field of crop taxonomies.

Relating to applications of HCAT for agricultural and ecological studies, literature shows that precise knowledge of the presence, variety, and

spatiotemporal organization of cultivated crops in a landscape can be a significant information source for elucidating ecological conditions and fluctuations on national and transnational levels. To outline technical considerations when working with HCAT and to demonstrate the usability of HCAT for crop-related diversity assessments, we conducted a case study comparing crop diversity at a landscape scale across seven EU regions. Having harmonized the national-specific datasets with HCAT, we first performed a comparison of crop diversity based on the Shannon index H_i and the number of crop types per spatial subunit $n_{c,i}$ (here: 1 km × 1 km grid cell). We could thereby observe intraregional and interregional patterns in crop diversity, which both decreased in heterogeneity with lower HCAT levels. On an intraregional scale, this is an expected behavior as the overall number of crop types in each region's taxonomy reduces with decreasing HCAT levels. On an interregional scale, this also holds true as the taxonomies increasingly align with each other with decreasing levels of detail in the crop-type denominations (see Fig. 3).

Taking the complete selection of regions without prior assessments allows, in any case, for studying qualitative trends between and within regions at varying HCAT levels. However, for a quantitative comparison of crop diversity across several regions, the scope of regions must be carefully selected with respect to the total number of crop types in their national declarations. As Supplementary Table 2 outlines, the number of crop types reported ranges from 15 in Croatia to 326 in the Netherlands. A comparison of the crop diversity as reported by the countries inevitably shows much lower crop diversity values for Croatia than for the Netherlands. Such differences can also be observed in the case study when comparing Lithuania (LI) with 24 classes compared to Flanders in Belgium (VLG) with 274 (Supplementary Figs. 1 and 2). There are two possible interpretations for these differences: (1) the diversity of cultivated crops is in reality higher for countries or regions such as Flanders in Belgium, or that (2) countries with a low number of crop types in their declaration taxonomies do not specify crop cultivations as detailed as others. Consequently, there would be a bias in the data when countries provide crop declarations only on a broader level, even though farmers might, in practice, cultivate more diverse sub-types.

Our case study provides evidence for both scenarios by documenting a variation in the similarity between regions when using crop types at higher HCAT levels (e.g., Lithuania compared to Austria (Fig. 4)) and also regions where the differences in crop diversity are apparent in all HCAT levels (e.g., Flanders and Slovakia). In the first case, differences in measures of crop diversity result from differences in the level of detail at which countries report the data and do not reflect real-world differences. In the latter case, differences in measures of crop diversity reflect differences in what farmers plant in the real world. Thus, when using these measures of crop diversity, e.g., as predictors in studies of biodiversity, we would expect meaningful relationships only in the latter case. Therefore, the similarity of crop declarations is an essential factor in the study design and evaluation process of studies relating, for instance, to biodiversity farming practices in studies that include areas of more than a single EU national state. Despite the Shannon index being defined both by the number of crop types per area and a factor describing spatial characteristics, interestingly, many regions (6 out of 8) strongly followed the patterns of the pure $n_{c,i}$ -based approach. Two countries (Austria and Slovenia) H_i indices were also dominated by spatial factors instead, showing deviating patterns compared to the other regions and the $n_{c,i}$ -based approach. These observations outline that knowledge of the presence of cultivated crop types is key in the context of crop diversity assessments, but they also confirm literature's reports on the importance of deriving landscape crop diversity together with its spatial characteristics^{6,16–18}. This further highlights the significance of joining HCAT with geodata, as it is established within the harmonized geodata set EuroCrops, which combines both spatial and harmonized crop data. HCAT served in this study as an accessible and user-friendly tool for analyzing crop management patterns in agricultural landscapes across various regions within the EU. These regions apparently exhibit distinct variations, which originate in practice from their characteristic socio-economic and geographic circumstances

and, at an administrative level, from a country's specific crop declaration system.

The presented case study concentrated solely on evaluating crop diversity, and on highlighting technical considerations when working with HCAT. We must emphasize, however, that with our chosen study design, there can not be drawn direct conclusions on biodiversity drivers without additional adjustments. This is mainly because we included all available crop types ranging from temporary, productive crops to permanent, non-cultivated vegetation types. Such an approach leads to an untraceable mix of crops regarding these two influential extra categories of cultivation intensity and retention period. Indeed, high values for our chosen crop diversity indices in one intensely cultivated area would indicate potentially more beneficial conditions for natural species compared to another area given the same level of cultivation intensity. Yet, areas with large, extensively cultivated, or even non-cultivated fields would show low crop diversity values but actually provide more beneficial conditions for biodiversity compared to highly crop-diverse but intensely managed fields. Consequently, for robust results, the selection of HCAT classes needs to be adjusted beforehand to a study's application-specific objective by, e.g., additionally considering biodiversity-related factors, including habitat structure, cultivation intensity, or food resources.

We gave an idea of HCAT's usage in the context of biodiversity research, yet HCAT's low level of customization also offers potential for application in various other domains. HCAT can serve as a flexible starting point towards thematic taxonomies using domain-specific extensions. These extensions may include thematic extra attributes, e.g., as in the FAO's ICC taxonomy³⁴, which would provide additional parameters for more customized application-oriented filtering or restructuring of the underlying data set. Taking the example of the three dimensions of sustainability, these extensions could range from economic masks that facilitate the evaluation of EU-wide food safety via social analyses on farm structures to tailored environmental studies on biodiversity considering the above-mentioned aspects. With the release of HCATv2, the EuroCrops project now increasingly intensifies communication and promotes the application of its results into various domains. Developing thematic taxonomies alone goes, however, beyond the current project team's capacities and expertise; EuroCrops, therefore, encourages each discipline to interact with the project team and the HCAT community for discussions and collective developments on domain-specific extension forms (e.g., via the HCAT-GitHub repository³⁶). The exemplary assessment of crop diversity represents a showcase of HCAT's applicability, we, however, want to highlight that the main result of our work is the taxonomy itself. Despite the mentioned limitations, this second version of HCAT offers significant potential for transnational analysis of spatiotemporal crop heterogeneity, diversity, and the structure of agricultural land, and by this, a meaningful source of information for transnational research on agricultural practices and related sustainability impacts.

Methods

Data collection

Within the scope of the EuroCrops project, data from the EU's Integrated Administration and Control System (IACS) was collected to compile a transnational dataset for crop type classification³⁵. Depending on a country's policies, it is possible to obtain geodata and related crop-type declarations via public websites or Web Feature Service (WFS); however, it was often necessary to contact the agricultural authorities directly. One integral property of the EuroCrops initiative is open and FAIR³⁷ access to all collections. We, therefore, ensured and communicated that all data given to us will be distributed under the Creative Commons Attribution International license CC BY-SA 4.0. After over one year of extensive outreach and communication, the resulting pool of gathered data exceeded our expectations: 13 countries and four subregions from three more countries contributed to the HCATv2 development by providing their geodata-based crop declarations including their national/regional crop taxonomies (compare Supplementary Table 2 and Fig. 2). The remaining EU members

states were still in process or denied to publicly share the data. In theory, having the possibility to use a large-scale unified dataset could build the foundation of an endless number of applications and use cases, especially in machine learning and data-driven modeling. However, publicly available crop data naturally comes in the national language of the respective country and uses country-dependent agricultural terms, leaving even common translation programs clueless. Supplementary Table 6 gives an insight into the original data and highlights the importance and need for a common crop taxonomy when designing transnational studies. The HCAT-enriched version of this original data is displayed in Supplementary Table 7, showing the additional HCAT-specific attributes comprising a translation, unique name, and code.

The development of HCATv2 required a volume of data that (a) aimed towards a large and geographically diverse as possible coverage of the studied area, i.e., the European Union, and (b) is lightweight enough for reasonably fast processing. We therefore decided to use only selected periods of the comprehensive multi-year and multinational EuroCrops database: from the period 2018–2021 we selected for each country the one, most recent year of IACS data being available at the time of data collection in mid-2021. Apart from the original situation of some countries having published their data for only a single, or few, non-consecutive years, there's also the aspect of some countries having released national datasets in retrospect with distinct delays (e.g., France). This is why the chosen base years for HCATv2 (a) can differ between countries/regions, and (b) may also deviate from the one that would be chosen based on the nowadays available database for the considered period. From this data collection, we subsequently extracted all available crop-type classes. The latter represents the set of data on which HCATv2 is based. Supplementary Table 2 displays the considered year for each country and gives an overview of the total number of crop parcels and occurring classes. The extensive list of each country's original crop types with their corresponding HCAT equivalents is publicly available on the EuroCrops' project GitHub repository³⁶. It is constantly updated with additional years and newly participating countries, serving as the basis for further extensions and improvements in future HCAT versions.

Transnational crop class harmonization

During the development of the EuroCrop project's first demo dataset²⁴, we were inspired by the already existing EAGLE classification scheme³² for the structuring of crop types. This scheme provides a broad hierarchical approach for the categorization of land use and land cover types ranging from natural biotic and abiotic to various anthropogenic land cover components. At the time of development, we found, however, a significant shortcoming in the EAGLE taxonomy regarding the scope and level of detail of crop types. We, therefore, already genuinely extended the EAGLE scheme by adding frequently occurring crop classes, leading to a first version of the Hierarchical Crop and Agriculture Taxonomy (HCATv1) as first presented in 2021²⁴.

Underestimating the demand for the proposed dataset and the variety of use cases, feedback taught us that the initially proposed scope of crop classes in HCATv1 needed more coverage and detail and made us rethink the entire process again. For HCATv2, we aimed to keep as much information as necessary while reducing the overhead as much as possible. In our case, this resulted in an iterative process of mapping almost all translated crop classes that we could obtain from collaborating countries and regions to our best knowledge to a dynamically growing extension of HCATv1. This time, we put a stronger focus on the hierarchical structure of the cultivated crop classes regarding land use, land cover, and biological families. Eventually, the number of classes increased from previously less than 100 to over 350. Despite this further extension of classes, there is still some inherent abstraction in the scope of crop types, as we summarized some rarely occurring, very similar, and/or single country-specific classes into related, overarching crop types. All original crop notations can, however, still be retraced using the provided original and translated name.

In the updated hierarchy, the sixth level entails the highest degree of detail. With decreasing numbers, the denominations for crop types become coarser (e.g., winter barley (level 6) \subset barley (level 5) \subset cereal (level 4) \subset arable crops (level 3)), with the third level representing the coarsest level that still differentiates between crop classes. HCAT levels 2 and 1 are given by the position of crop types within the EAGLE taxonomy version 1³². In this EAGLE version, crop types are classified into the category "Crop Type" (level 2), which itself is part of the category "Land Characteristics" (level 1), both being numbered by the digit 3 in their respective hierarchy layer. Consequently, we defined each of the HCAT codes to begin with the prefix "33" (see column EC_hcat_c of Supplementary Table 7), indicating this placement into the higher-level land use and land cover context.

Case study: cultivated crop diversity in Europe

To showcase the practical application of HCAT in research on crop-related diversity, we designed a case study focusing on assessing crop diversity across seven EU regions at a landscape scale. The specific subgoals of this study were to identify distinctive patterns (1) in relation to different HCAT levels and (2) between each other. Moreover, the study serves to highlight key factors that should be considered when conducting crop diversity assessments using HCAT. The initial selection of countries/regions was influenced by the public availability of national data sets and the intention to include representatives from various areas of the EU, including countries from the Atlantic region with the German federal state of North Rhine-Westphalia and Flanders in Belgium, the Continental region with Austria and North Rhine-Westphalia, the Boreal region with Lithuania and Latvia, Alpine areas with Austria and Slovenia, and Pannonian region with Slovakia. Exhibiting limited crop reporting at only HCAT level 4, we could not incorporate the regions/countries Navarra, Croatia, and Romania, which otherwise would geographically depict suitable representatives for the Mediterranean and Steppic regions. The same holds for Portugal, which shared its national taxonomy with us but has not provided the necessary geodata so far. The mentioned regional classes originated from the Biogeographic Regions classification scheme of the European Environment Agency. Areas within these regions exhibit similarities encompassing factors such as climatic conditions, geological attributes, and vegetation patterns³⁸.

In the first step, we harmonized each region-specific EuroCrops-sub dataset of original crop declarations by extending these with the HCAT crop declaration. To compare across several HCAT levels, we created three duplicates per region—each with a different HCAT level ranging from 4 to 6. After this step, a country's overarching, application-specific filtering of crop types would be seamlessly possible. For our application independent calculation examples, we took the entire scope of available crop types, including temporary crops, but also permanent cultivated and non-cultivated types. Next, in order to have a common spatial resolution for transregional/national statistical comparisons, we split each region-specific dataset into grids of 1 km \times 1 km cells. This process was done by first constructing vector-based grids that span the entire area of each respective region; afterward, we intersected each grid cell with the vector-based parcel data. After these pre-processing steps, we iterated for each of the 24 combinations of scales (8 regions times 3 HCAT levels) over all the grid cells and derived statistical metrics, including the number of parcels in a grid cell, the contained area of each parcel, and the list of cultivated crop types for each cell. We chose the number of crop types in a grid cell $n_{c,i}$ as one measure for exploring each region's crop diversity across the three different HCAT levels. $n_{c,i}$ represents a metric that allows for quick qualitative estimation of crop diversity without additional calculations. It, however, neglects both number and area-wise distributions of samples inside each cell's population of crop parcels, which are likewise influential factors to crop diversity at landscape scale (see Introduction). Relying only on $n_{c,i}$ might, therefore, lead to insufficient results during quantitative diversity assessments, which is why we decided to also calculate the Shannon index H_i for each grid cell i . H_i represents a common diversity index, which describes, in our case, the

<https://doi.org/10.1038/s44264-024-00037-x>

Article

diversity of crop types, considering both the number of crop types and their abundance, expressed here through the relative area in the grid cell³⁹. It is defined as

$$H_i = - \sum_k p_{k,i} \cdot \ln(p_{k,i}) \quad (1)$$

with

$$p_{k,i} = \frac{n_{C_i,k}}{N_{C_i}} \quad (2)$$

where $p_{k,i}$ denotes the ratio of each crop type k 's area $n_{C_i,k}$ in grid cell i and the total area of samples N_{C_i} in grid cell i . To answer the described study objectives, we calculated descriptive statistics for each region's $n_{c,i}$, H_i , as well as the number of parcels and the median field area per grid cell as potential indicators for spatial variability. We applied a common color scale to each metric's set of maps to reveal the quantitative differences between considered countries.

Data availability

One core aspect of the EuroCrops project is the access and distribution of open, freely available, but sometimes hidden data. Allowing everyone to work with and take part in collecting the dataset benefits society in a much larger sense than encapsulating the gained knowledge and actively restricting data access. EuroCrops, the HCAT taxonomy, and all mappings from the country-specific data can therefore be found on our website⁴⁰, GitHub repository⁴⁶ and several platforms such as EO-Lab⁴¹, GeoDB⁴² and soon on the AWS Open Data Sponsorship Program⁴³. All datasets we distribute are licensed under CC BY-SA 4.0, and everything published in the future will follow this approach, ensuring a taxonomy, mapping, and dataset that follow the FAIR principles³⁷. The datasets analyzed during the current work are available in the EuroCrops Zenodo repository⁴⁴.

Code availability

The code used in this research is not publicly available. However, the methods and analyses described in the article are based on standard data science and geospatial data libraries for Python 3, including Pandas, NumPy, Matplotlib, GeoPandas, and Shapely. Researchers interested in replicating the study can do so by using these widely available libraries and following the methodological details provided in the paper.

Received: 10 January 2024; Accepted: 20 November 2024;

Published online: 20 January 2025

References

1. Food and Agriculture Organization of the United Nations. Natural Capital Impacts in Agriculture. Supporting better business decision-making (2015).
2. Masson-Delmotte, V. et al. (eds.). *Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change* (Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, 2021).
3. Burns, F. et al. Abundance decline in the avifauna of the European Union reveals cross-continental similarities in biodiversity change. *Ecol. Evol.* **11**, 16647–16660 (2021).
4. Rosenberg, K. V. et al. Decline of the North American avifauna. *Science* **366**, 120–124 (2019).
5. Chase, J. M., Blowes, S. A., Knight, T. M., Gerstner, K. & May, F. Ecosystem decay exacerbates biodiversity loss with habitat loss. *Nature* **584**, 238–243 (2020).
6. Clough, Y., Kirchweber, S. & Kantelhardt, J. Field sizes and the future of farmland biodiversity in European landscapes. *Conserv. Lett.* **13**, e12752 (2020).
7. Benton, T. G., Vickery, J. A. & Wilson, J. D. Farmland biodiversity: is habitat heterogeneity the key? *Trends Ecol. Evol.* **18**, 182–188 (2003).
8. Skaloš, J., Molnárová, K. & Kottová, P. Land reforms reflected in the farming landscape in East Bohemia and in Southern Sweden—two faces of modernisation. *Appl. Geogr.* **35**, 114–123 (2012).
9. Tryjanowski, P. et al. Conservation of farmland birds faces different challenges in Western and Central-Eastern Europe. *Acta Ornithol.* **46**, 1–12 (2011).
10. Brooks, D. R. et al. Large carabid beetle declines in a United Kingdom monitoring network increases evidence for a widespread loss in insect biodiversity. *J. Appl. Ecol.* **49**, 1009–1019 (2012).
11. Gregory, R. D., Skorpilova, J., Vorisek, P. & Butler, S. An analysis of trends, uncertainty and species selection shows contrasting trends of widespread forest and farmland birds in Europe. *Ecol. Indic.* **103**, 676–687 (2019).
12. Seibold, S. et al. Arthropod decline in grasslands and forests is associated with landscape-level drivers. *Nature* **574**, 671–674 (2019).
13. Holzschuh, A. et al. Mass-flowering crops dilute pollinator abundance in agricultural landscapes across Europe. *Ecol. Lett.* **19**, 1228–1236 (2016).
14. Rundlöf, M., Persson, A. S., Smith, H. G. & Bommarco, R. Late-season mass-flowering red clover increases bumble bee queen and male densities. *Biol. Conserv.* **172**, 138–145 (2014).
15. Palmu, E., Ekroos, J., Hanson, H. I., Smith, H. G. & Hedlund, K. Landscape-scale crop diversity interacts with local management to determine ground beetle diversity. *Basic Appl. Ecol.* **15**, 241–249 (2014).
16. Aguilera, G. et al. Crop diversity benefits carabid and pollinator communities in landscapes with semi-natural habitats. *J. Appl. Ecol.* **57**, 2170–2179 (2020).
17. Martin, E. A. et al. The interplay of landscape composition and configuration: new pathways to manage functional biodiversity and agroecosystem services across Europe. *Ecol. Lett.* **22**, 1083–1094 (2019).
18. Sirami, C. et al. Increasing crop heterogeneity enhances multitrophic diversity across agricultural regions. *Proc. Natl. Acad. Sci. USA* **116**, 16442–16447 (2019).
19. Billeter, R. et al. Indicators for biodiversity in agricultural landscapes: a pan-European study. *J. Appl. Ecol.* **45**, 141–150 (2008).
20. European Commission. *EU Biodiversity Strategy for 2030. Bringing Nature Back Into Our Lives*. 1st ed. (Publications Office of the European Union, Luxembourg, 2021).
21. Moersberger, H. et al. *Europa Biodiversity Observation Network: User and Policy Needs Assessment* (ARPHA Preprints, 2022).
22. IPBES. Global assessment report on biodiversity and ecosystem services of the Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services (2019).
23. Fall, J. J. Conservation across borders: biodiversity in an interdependent world. *Mt. Res. Dev.* **29**, 103 (2009).
24. Schneider, M., Broszeit, A. & Körner, M. EuroCrops: a Pan-European dataset for time series crop type classification. Preprint at *arXiv* <https://doi.org/10.48550/arXiv.2106.08151> (2021).
25. Turkoglu, M. O. et al. Crop mapping from image time series: deep learning with multi-scale label hierarchies. *Remote Sens. Environ.* **264**, 112603 (2021).
26. Rußwurm, M., Pelletier, C., Zollner, M., Lefèvre, S. & Körner, M. BreizhCrops: a time series dataset for crop type mapping. *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.*, XLIII-B2-2020, 1545–1551, <https://doi.org/10.5194/isprs-archives-XLIII-B2-2020-1545-2020> (2020).
27. European Union. *Regulation (EU) No 1307/2013 of the European Parliament and of the Council of 17 December 2013 establishing rules for direct payments to farmers under support schemes within the framework of the common agricultural policy and repealing Council Regulation (EC) No 637/2008 and Council Regulation (EC) No 73/2009* (2013).

28. European Union. *Regulation (EU) 2021/2115 of the European Parliament and of the Council of 2 December 2021 establishing rules on support for strategic plans to be drawn up by Member States under the common agricultural policy (CAP Strategic Plans) and financed by the European Agricultural Guarantee Fund (EAGF) and by the European Agricultural Fund for Rural Development (EAFRD) and repealing Regulations (EU) No 1305/2013 and (EU) No 1307/2013* (2023).
29. European Union. *Commission Implementing Regulation (EU) No 809/2014 of 17 July 2014 laying down rules for the application of Regulation (EU) No 1306/2013 of the European Parliament and of the Council with regard to the integrated administration and control system, rural development measures and cross compliance* (2014).
30. Gamot, V. S. F., Landrieu, L., Giordano, S. & Chehata, N. Satellite Image Time Series Classification with Pixel-Set Encoders and Temporal Self-Attention. 18.11.2019.
31. Rußwurm, M. & Körner, M. Multi-temporal land cover classification with long short-term memory neural networks. *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.* <https://doi.org/10.5194/isprs-archives-XLII-1-W1-551-2017> (2017).
32. Arnold, S. et al. The EAGLE Concept—A Vision of a Future European Land Monitoring Framework (2013).
33. Büttner, G. CORINE land cover and land cover change products. In *Land Use and Land Cover Mapping in Europe*. (eds I. Manakos & M. Braun) (Springer Netherlands, Dordrecht, 2014), 18, pp. 55–74.
34. Food and Agriculture Organization of the United Nations. *World Programme for the Census of Agriculture 2020* (Food and Agriculture Organization of the United Nations, Rome, 2018).
35. Schneider, M., Schelte, T., Schmitz, F. & Körner, M. EuroCrops: the largest harmonized open crop dataset across the European Union. *Sci. Data* **10**, 612 (2023).
36. Schneider, M. EuroCrops. *GitHub Repository*. <https://github.com/maja601/EuroCrops> (2023).
37. Wilkinson, M. D. et al. The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data* **3**, 160018 (2016).
38. Cervellini, M. et al. A grid-based map for the Biogeographical Regions of Europe. *Biodivers. Data J.* **8**, e53720 (2020).
39. Krebs, C. J. *Ecological Methodology* (Addison Wesley Longman, Menlo Park, California, 2007).
40. Schneider, M. *EuroCrops*. <https://www.eurocrops.tum.de/> (2023).
41. German Aerospace Center. *EOlab Portfolio*. <https://eo-lab.org/en/portfolio/> (2023).
42. Euro Data Cube Consortium. *Sentinel-2 Signals for EuroCrops*. <https://collections.eurodatacube.com/sentinel-2-signals-for-eurocrops/> (2023).
43. Amazon Web Services, Inc. *Open Data Sponsorship-Programm*. <https://aws.amazon.com/de/opensource/open-data-sponsorship-program/> (2023).
44. Schneider, M., Chan, A. & Körner, M. EuroCrops. *Zenodo Data Repository*. <https://zenodo.org/records/10118572> (2023).

Acknowledgements

The authors and the EuroCrops project receive funding from the German Federal Ministry for Economic Affairs and Climate Action on the basis of a resolution of the German Bundestag under reference 50EE1908 and from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101004112.

Author contributions

M.S. and M.K. lead the EuroCrops project, where M.S. collected the data and developed the taxonomy. D.G., M.S. and S.M. developed the concept for introducing the taxonomy into the agricultural and ecological context, including the discussion on HCATv2's implications for practical application. D.G. and M.S. developed the experimental design. M.S. and D.G. prepared the visualizations and tables. Together with D.G., J.P. developed the software necessary for the case study. All authors contributed to writing the paper.

Funding

Open Access funding enabled and organized by Projekt DEAL.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s44264-024-00037-x>.

Correspondence and requests for materials should be addressed to David Gackstetter.

Reprints and permissions information is available at <http://www.nature.com/reprints>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2025