Data-Centric Machine Learning for Geospatial Remote Sensing Data

Ribana Roscher[a,b], Marc Rußwurm[c], Caroline Gevaert[d], Michael Kampffmeyer[e], Jefersson A. dos Santos[f], Maria Vakalopoulou[i], Ronny Hänsch[h], Stine Hansen[e], Keiller Nogueira[i], Jonathan Prexl[j], Devis Tuia[k]

[a]*Data Science for Crop Systems Group, Forschungszentrum Jülich GmbH, Wilhelm-Johnen-Straße, Jülich, 52428, Germany*
[b]*Remote Sensing Group, University of Bonn, Niebuhrstr. 1a, Bonn, 53113, Germany*
[c]*Laboratory of Geo-information Science and Remote Sensing, Wageningen University, Droevendaalsesteeg 3, Wageningen, 6708 PB, Gelderland, Netherlands*
[d]*Department of Earth Observation Science, Faculty ITC, University of Twente, Drienerlolaan 5, Enschede, 7522 NB, Overijssel, the Netherlands*
[e]*Department of Physics and Technology, UiT The Arctic University of Norway, Klokkargårdsbakken 35, Tromsø, 9019, Norway*
[f]*Department of Computer Science, University of Sheffield, 211 Portobello, Sheffield City Centre, Sheffield, S1 4DP, United Kingdom*
[g]*MICS Laboratory, CentraleSupélec, Paris-Saclay University, 3 Rue Joliot Curie, Gif-sur-Yvette, 91190, France*
[h]*Microwaves and Radar Institute, German Aerospace Center (DLR), Münchner Str. 20, Oberpfaffenhofen-Wessling, 82234, Germany*
[i]*Computing Science and Mathematics, University of Stirling, Stirling, FK9 4LA, United Kingdom*
[j]*Department of Aerospace Engineering, University of the Bundeswehr Munich, Werner-Heisenberg-Weg 39, Neubiberg, 85579, Bavaria, Germany*
[k]*Environmental Computational Science and Earth Observation Laboratory, Ecole Polytechnique Fédérale de Lausanne (EPFL), Ronquos 86, Sion, 1951, Switzerland*

Abstract

Recent developments and research in modern machine learning have led to substantial improvements in the geospatial field. Although numerous deep learning architectures and models have been proposed, the majority of them have been solely developed on benchmark datasets that lack strong real-world relevance. Furthermore, the performance of many methods has already saturated on these datasets. We argue that a shift from a model-centric view to a complementary data-centric perspective is necessary for further improvements in accuracy, generalization ability, and real impact on end-user appli-

cations. Furthermore, considering the entire machine learning cycle—from problem definition to model deployment with feedback—is crucial for enhancing machine learning models that can be reliable in unforeseen situations. This work presents a definition as well as a precise categorization and overview of automated data-centric learning approaches for geospatial data. It highlights the complementary role of data-centric learning with respect to model-centric in the larger machine learning deployment cycle. We review papers across the entire geospatial field and categorize them into different groups. A set of representative experiments shows concrete implementation examples. These examples provide concrete steps to act on geospatial data with data-centric machine learning approaches.

1. Introduction

Remote sensing data is a central link between pressing global challenges and the possibilities of machine learning methods. To realize its full potential, a deep understanding of the data and its utilization possibilities is essential. Generally, data plays a fundamental role in the machine learning (ML) cycle with steps from informing (1) problem definition, (2) data creation, (3) data curation, enabling (4) model training, and (5) evaluation to (6) the eventual model deployment that feeds back to modifying the problem definition. We show this cycle in Fig. 1 where each node is colored by its focus of being problem-centric (dark gray), data-centric (blue), or model-centric (dark green). Yet, current research in machine learning is predominantly model-centric and focuses on model design and evaluation (Step 4). This primarily emphasizes optimizing the accuracy and efficiency of the models themselves [1, 2, 3] and considers the dataset rather as a static benchmark than a dynamic representation of the application. Even in applied machine learning areas such as geospatial data analysis, the research towards integrating new machine learning methods has recently largely focused on refining the algorithms and fine-tuning the model parameters [4]. Data-centric aspects like data acquisition (Step 1) and curation (Step 2) are done manually to a large degree and data quality is rarely considered. At the same time, most domain expertise remains concentrated in classic feature engineering.
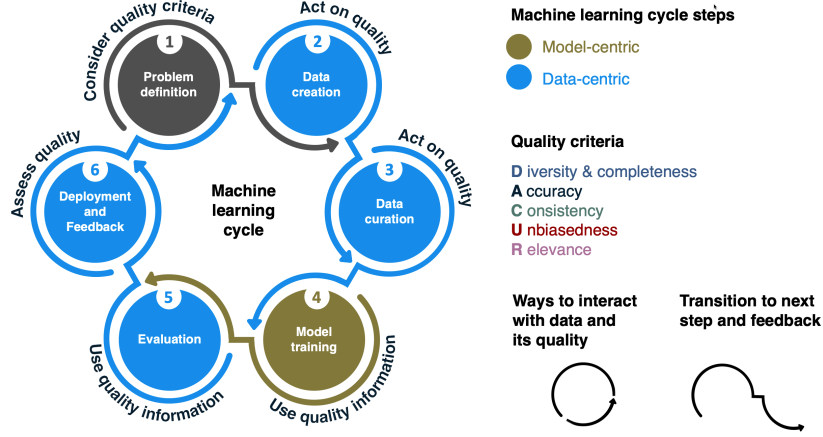
Figure 1: Steps of the machine learning cycle. Each step highlights one way to interact with data and its quality, and each step can employ multiple techniques to perform the interaction (see Section 3). It involves problem definition, data creation and curation, model training and evaluation, and final deployment, which feeds back into a modified problem definition. Model-centric learning focuses primarily on model training and evaluation, while data-centric learning involves algorithms covering data curation, creation, specific training strategies, evaluation, and deployment feedback. The considered quality criteria are diversity and completeness, accuracy, consistency, unbiasedness, and relevance, see Section 2).

The rapid development of machine learning methods has been primarily driven by the availability of large-scale datasets and advancements in computing power [5, 6]. This has allowed researchers to focus on developing complex models capable of capturing patterns and relationships within the data. However, many developments have been pursued in controlled benchmark settings only, where the natural variability and real-world characteristics of the data are bypassed. In the geospatial domain, for example, this is achieved through a one-time pre-defined data location sampling, as seen in the predefined regions of the Sen12MS dataset [7], or through class-balancing techniques commonly used in crop-type mapping [8]. Although the great importance of these benchmarks should not be diminished, as a result, many methods developed are often detached from reality and seldom robust in real-world deployment. This requires considering all steps in the machine learning pipeline and acknowledging it as a cycle (Fig. 1), as well as a high degree of robustness for acquisition specification and generalization to diverse shifts.

Despite the availability of enormous amounts of data [6], including well-curated benchmarks, and significant progress made through methodological advances, there is currently a saturation point where many established architectures and training methods achieve comparable accuracy ([9], 2023 AI Index Report[1]). This is exemplified by foundation models or general-purpose computing architectures for multiple modalities and tasks [10, 11, 12, 13]. These large-scale models, trained on massive but mainly domain-agnostic datasets, increasingly replace modality-specific recurrent and convolutional models. This development comes with the prevailing belief that "more data is better". However, the reality is more complex. While larger datasets can provide a broader representation of real-world scenarios, foundation models require adaptation through downstream models tailored for specific applications through fine-tuning datasets that can be substantially smaller in size. Still, the smaller the size, the more important the data quality becomes, as noisy labels and other errors can have increasingly detrimental effects on final accuracy. In this direction, it has already been argued that in addition to the quantity of data, the quality is a significant factor in model performance across various domains [14, 15, 16, 1, 17, 18]. That means that not all data samples contribute equally to model learning and generalization. In fact, indiscriminately increasing data volume can lead to diminishing returns or even detrimental effects on model performance. The assumption that more data inherently leads to better outcomes overlooks the complexities of data distribution, the potential for introducing biases, spurious correlations, and the computational resources required for processing and storing vast datasets [19, 20]. Moreover, many scientific disciplines, particularly those reliant on experimental data, face significant data scarcity. In these domains, the emphasis is on quality rather than quantity, as each data point is meticulously collected and holds substantial scientific value. Therefore, while foundation models represent a leap forward in AI capabilities, the general belief that "more data is better data" does not fully address the complexities of real-world applications. Effective solutions require an approach that considers both the quality and quantity of data, as well as the adaptability of models to specific tasks and domains.

However, a number of challenges need to be overcome before this can be turned into actions. The notion of data quality and its specific crite-

---

[1]https://aiindex.stanford.edu/report/

ria is only vaguely defined yet and can differ depending on the domain and task. In the geospatial domain, we have identified five criteria especially relevant to machine learning: diversity and completeness, accuracy, consistency, unbiasedness, and task-relevance, which will be discussed in this paper. Automating to interact with the data regarding their quality in each step in the machine learning cycle is crucial. This is especially important when dealing with training and evaluation datasets that are too large for traditional visual refinement and manual data manipulation. The importance of shifting to a more data-centric perspective is also underlined by the growing demand for methods that are robust to noise [21, 22] or missing data [23, 24, 25], or that are specifically designed to handle data quality deficiencies. Here, it becomes clear that assessing and optimizing data quality is crucial in addressing real-world data challenges. However, such methods represent only one way to interact with data regarding their quality (Step 4: use quality information during model training).

This article offers a systematic discussion of data-relevant aspects of the machine learning cycle in the context of geospatial data. It has the goal of fostering awareness about the benefits of optimizing geospatial data alongside new methods of research. Within the scope of this paper, we use the term

> **data-centric learning** as a paradigm focusing on the systematic, automated, and algorithmic determination, as well as the utilization of a rich and high-quality dataset, including a rigorous evaluation process to ensure that the model performs optimally on the dataset for the intended task.

In contrast, model-centric learning focuses on the optimization of the model parameters that include trainable parameters and hyperparameters, such as the model design, as well as the learning objective as a loss function. It also focuses on the definition of evaluation metrics to measure the model's performance with respect to a static evaluation dataset.

The article is organized as follows: It first clarifies the terminology, definitions, and goals of data-centric and model-centric learning (Section 2) and then provides a comprehensive discussion of data-centric machine learning and its steps, with a focus on the analysis and interpretation of geospatial data. Additionally, it presents a systematic review of data-centric machine learning techniques tailored explicitly to geospatial data (Section 3), facilitat-

5

ing the learning and evaluation of high-performing machine learning models (see Fig. 1). The paper is supported by three experiments using data-centric machine learning techniques, presented in Section 4.

2. Definitions and Terminology

This section describes the machine learning cycle, provides definitions of data-centric and model-centric learning, and details the pursued data quality criteria. We refer to data as a set of individual raw samples and their targets, where a data sample may be assigned to one or more targets.

2.1. Machine learning cycle

The machine learning cycle, illustrated in Fig. 1, involves the following sequential steps that may be iterated following the entire circle or within intermediate steps:

1. **Problem definition.** Definition of the underlying objectives by leveraging problem-centric expertise through domain knowledge. Methodologically, this step defines the ML task, e.g., classification or regression, the method, data type and sources based on the application's goals, needs, and restrictions.

2. **Data creation.** Generation or collection of data from various sources. This includes manually collecting data through surveys or measuring instruments, automatically acquiring sensor data, systematically and algorithmically scraping data from the Web, or generating data using generative models.

3. **Data curation.** Utilizing curation strategies to obtain high-quality data and prepare data for modeling. This can be done through outlier filtering, label noise identification and reduction, or de-biasing data with respect to relevance to, for instance, the region of interest.

4. **Model training and hyperparameter tuning.** Learn and fine-tune a model utilizing the curated data for training and validation. This involves choosing the model architecture and objective function.

5. **Evaluation.** Applying the machine learning model to evaluation data to assess model performance. This involves, in close connection to Steps 1 and 2, the creation and curation of evaluation data or defining a suitable test split that may be spatially distant from training areas.

6

6. **Model deployment, monitoring, and feedback.** Continuously monitoring the trained model regarding its performance in previously unknown real-world applications and giving feedback to improve the data and the model. This step generates feedback from the model deployment and the monitoring to inform about changed needs and goals and to improve the overall process.

Within the cycle, data-centric learning focuses on Steps 2: *Curation*, 3: *Creation*, 5: *Evaluation*, and 6: *Deployment*. In contrast, model-centric learning involves Steps 4: *Model training* and 5: *Evaluation* in terms of metric definitions.

2.2. High-quality criteria for geospatial data

While model-centric learning approaches focus on the optimization of performance on a static evaluation set, data-centric approaches focus on a series of quality criteria of the data used. We identified five criteria that are defined in the following. We further indicate whether each criterion is typically considered at a global (dataset-wide) level or a local (individual data sample or subset) level. Please note that some criteria have already been pointed out or other criteria might be more important in specific applications (e.g., [26]). Moreover, criteria can be closely connected or partly overlap in specific settings, however, we consider them as distinct concepts that describe different characteristics of the data. This is demonstrated, for example, by the fact that if one quality criterion is high, other criteria do not automatically have to be high as well. We consider the data to be observations and labels and do not make an explicit distinction here.

- **Diversity and completeness (global).** Completeness refers to the extent to which all necessary data samples are present within the dataset. It involves ensuring that there are no missing data points and that the dataset includes all relevant information required for the specific task. In the geospatial domain, completeness is critical as missing data can arise due to sparse data acquisition, multimodal fusion with different resolutions, or challenges in gathering information from different sensors. Techniques to measure completeness include data coverage analysis and gap analysis, while metrics can include the percentage of missing values and the spatial coverage ratio. Closely related, diversity pertains to the inclusion of a wide range of data samples that represent

various scenarios and variations, enabling ML models to handle various real-world scenarios [27]. In the context of geospatial data, this means including data from various geographic areas, different times, and multiple sensor types to capture the full range of potential conditions. It is important to consider diversity within the context of completeness, which ensures that all necessary information is covered without important details missing.

- **Accuracy (local).** Data accuracy indicates how closely the data values align with the true or expected values. It quantifies the correctness and precision of the data, accounting for any errors or deviations that might occur during data collection. In the geospatial domain, accuracy is influenced by factors such as random variations or errors during data acquisition, sensor precision, environmental conditions, and data processing techniques [28]. Additionally, human errors in generating reference labels [29], or situations where the phenomenon targeted by machine learning models is conceptually ambiguous, such as identifying informal settlements from satellite imagery [30] or classifying wilderness [31, 32], contribute to uncertainty. Accuracy can be, for example, assessed by repeated measures or comparing the data against reference values.

- **Consistency (local and global).** Consistency, locally, refers to the uniformity of properties or attributes of single data samples within one dataset and across multiple datasets. Globally, it ensures that the dataset maintains uniformity in terms of data formatting, units, and definitions across all samples. For geospatial data, consistency is crucial when integrating data from different sources, such as varying coordinate systems or temporal resolutions. In the case of machine learning tasks for geographic data, ensuring consistency is critical during the process of annotating samples for training models [33, 34].

- **Unbiasedness (global)**. Unbiasedness measures the extent to which data is free from systematic errors and and consistent deviations. An unbiased dataset accurately reflects the true values and distributions without any systematic over- or underestimation. In the geospatial domain, biases can be manifold [14]. For example, acquiring data is generally easier in data-rich geographic areas such as Europe or North America. However, it can be much more challenging in other continents with limited data availability, which can introduce bias [14, 27].

- **Relevance (local).** Relevance denotes how applicable and suitable

a data sample is for the specific task or problem being addressed. It assesses the alignment of the data with the requirements and objectives of the analysis or model. For geospatial data, relevance involves ensuring the data is up-to-date and reflect the present conditions of the Earth. As already pointed out by [2], datasets and models need to be updated continuously to be relevant.

Generally, having measures for the data quality allows quantifying the goodness of a single sample (locally) or the whole dataset (globally) with respect to a given criterion, which provides the user with intuition and builds the basis for actions in data-centric approaches. A measure also helps identify issues in the dataset that would hinder the development of a well-performing model. Once the quality is quantified, specific steps can be taken to address the identified issues and modify the dataset. Quality measures serve as a guide for creating, collecting, enhancing, and curating the data. Additionally, a quantitative measure enables us to prioritize the steps involved. Existing approaches to quantify the quality (value) of data can be divided into label-based [35, 36] and label-free approaches [37]. Label-based approaches typically evaluate the effect that a given sample has on the validation loss. In contrast, the label-free approaches leverage inherent data characteristics to assign values according to the desired quality criteria. While a quantification of quality can find direct applications in the form of data marketing [38], it more often is used to facilitate downstream tasks such as active learning [39] or data cleaning [35], where the value is used to select relevant or prune irrelevant data samples.

3. Data-centric machine learning techniques

In this section, we thoroughly review the literature and present data-centric techniques for the machine learning Steps 2-5 with a specific focus on geospatial remote sensing data. We skip Step 1, since it mainly contains manually performed actions, and we skip Step 6 since it mainly is concerned with the feedback loop. We group and structure these techniques in Fig. 2 with respect to the mentioned quality criteria. Each step in the machine learning cycle interacts with data in a specific way regarding its quality. This includes acting on data quality, using quality information, assessing quality, and considering the quality. The review aims at consolidating knowledge helping to quickly understand the current state of the field, building the basis for dentifying gaps and opportunities for future research.
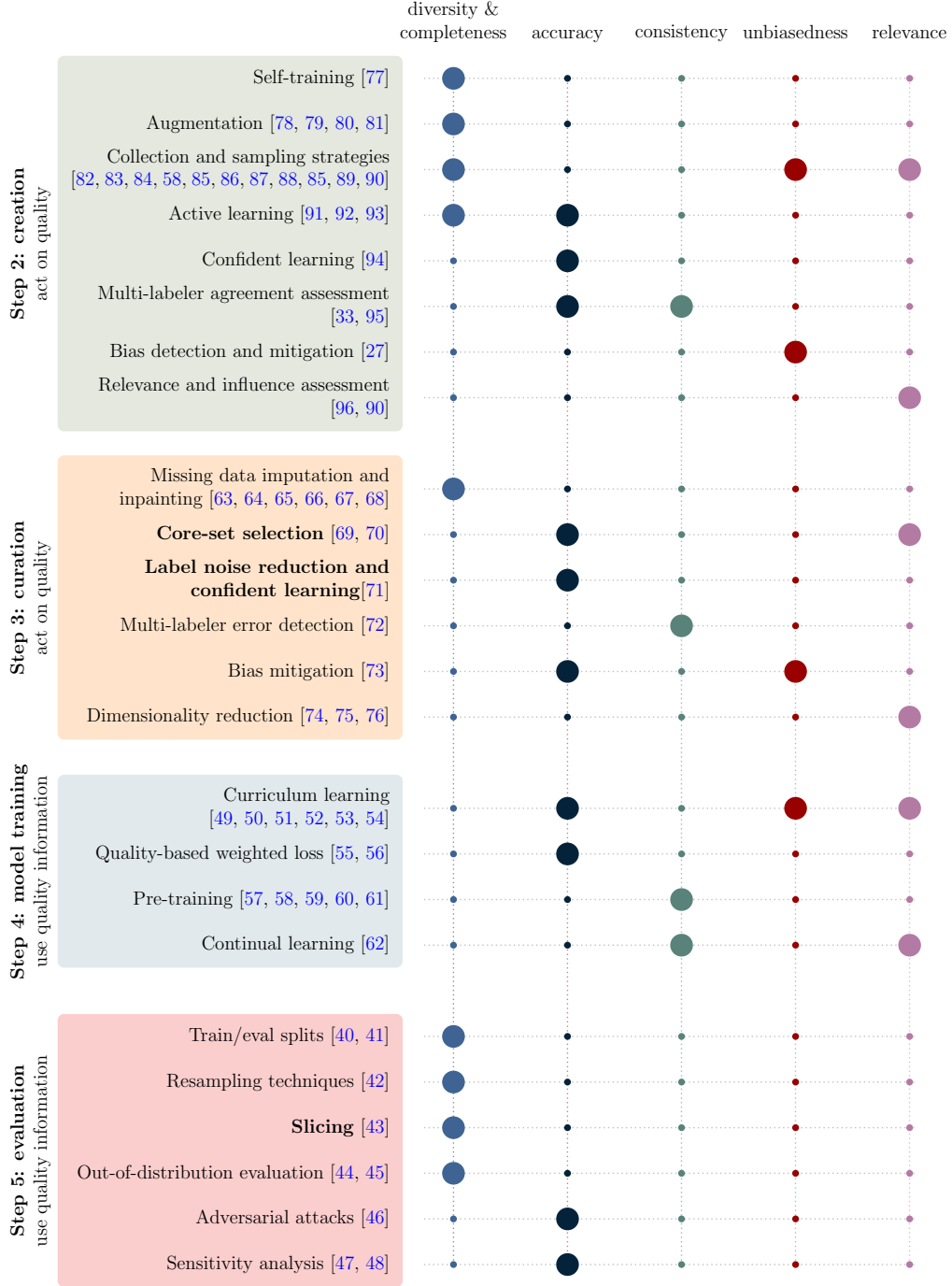
9

Figure 2: Data-centric machine learning techniques and papers referenced in Section 3. Each machine learning step and each technique (rows) interacts in a specific way with the data and the quality (columns). The techniques used for our experiments are highlighted in bold. Large dots indicate which quality information is used (model training and evaluation), or which specific quality criteria is acted on (in the creation and curation step).

3.1. Data Creation (Step 2)

Data creation, whether through collection or generation, directly acts on the data quality. It serves two distinct purposes: firstly, high-quality training data supports the model learning process, while secondly, high-quality evaluation data enhances the information content to provide deeper insights into the model's capabilities (see also Step 5: *Model evaluation*).

Diversity and completeness. Fostering diversity and completeness are closely related to the sampling scheme, which should ensure heterogeneity, such as across different geographic regions (spatial diversity) or related environmental conditions, for example, to seasonal variations (temporal diversity). Measuring the heterogeneity of samples in geospatial applications can be based, for example, on spectral similarity [86, 87] or morphological image characteristics [88]. Wesley et al. [97] propose the Geospatial Inception Score, which considers diversity as a function of the co-occurrence of samples, their variety, and their context. Another measure is presented by Xu et al. [37], who propose a quantification of diversity in terms of the volume of the data matrix (determinant of its left Gram matrix). A sample's contribution to diversity can then be measured via the increase in volume when adding it to the existing dataset. Generally, sampling schemes should be designed to cover the expected diversity of the classes and to avoid biases. For example, stratified random sampling is a long-standing good practice in geospatial applications [82, 83]. Further efforts to improve diversity and completeness involve techniques to merge data from multiple sensors (e.g., optical, radar, LiDAR, or thermal) and collecting data at different spatial scales to create more comprehensive datasets [85, 89, 85, 84]. This ensures that data encompasses a wide range of features, supporting an understanding of the geographical area under study. However, research in this direction has so far been the exception rather than the rule and a close interaction with model-centric developments is necessary to find suitable strategies to integrate multi-resolution, multi-temporal, and multi-spectral data in a joint framework [14, 98].

Moreover, it is not always feasible to collect a sufficient amount of data. This is particularly the case if the reference data is collected via in-situ measurements which is often time-consuming and a high effort [99]. To address this limitation, data augmentation has become a prominent data-centric technique. This technique involves artificially increasing the size of the dataset by applying various geometric and spectral transformations to the input data samples [78]. Also, generative adversarial networks (GANs) and diffusion

11

models are increasingly used to generate synthetic data samples and augment the existing data. This helps to diversify the dataset by creating samples coherent with the original data distribution [80, 81, 79]. Although the general concepts are promising, research gaps comprise the kind of augmentation to ensure a physically consistent model and the capturing of powerful features from the data [100, 101].

If a learned model is already available, it can be used to acquire new data through self-training [77]. In self-training, the trained model is used to classify unlabeled samples, and the predictions are used as pseudo-labels. Samples with their pseudo-labels are then converted into labeled training data based on specific quality criteria. Another technique is active learning [91]. Here, the focus is on identifying unlabeled data samples that would be most useful and should be added to an existing dataset. The labeling is performed manually, a process known as human-in-the-loop. Used metrics in this context are the Most Ambiguous and Orthogonal (MAO) metric [93] and angle-based diversity [92] that utilizes the angle between two samples in a feature space to measure diversity. However, the selection of samples is still an open research question [102] and current research point towards a necessity to use multiple quality criteria beyond diversity and completeness, so far mainly focusing on uncertainty.

Accuracy. Fang et al. [94] employ confident learning [103] to select high-confident training samples in a domain adaptation setting from a set of pseudo-labeled samples in the target domain scene. The confidence of the labels is assessed by comparing their reference label with the estimated score for that label. A remaining challenge, however, is the notion of data accuracy and confidence and their disentanglement from model uncertainty, with uncertainty quantification being a research area gaining increased attention [28]. Bastani et al. [104] conduct iterative analyses during the design of their SATLASPRETRAIN dataset to evaluate the precision and recall of labels collected from various sources (e.g., manually labeled and from existing products). This process helps improve the overall quality of the labels. Additionally, they provide information on the accuracy of the categories determined during dataset creation. With knowledge about the problem-specific patterns, label refinement strategies can be employed, for example, by improving coarse hand-annotations. Here, for instance, Rußwurm et al. [105] designed a simple computer vision pipeline to refine coarse hand-labels of floating plastic marine debris from Sentinel-2 images. This label refinement improved the overall model accuracy of the segmentation model.

12

Consistency. Data consistency is maintained by implementing rigorous data collection procedures, which include standardizing measurements or labeling procedures, and quality control measures. Ensuring data consistency can be particularly challenging when multiple experts label reference data, such as through crowdsourcing platforms [106]. While online crowdsourcing platforms like OpenStreetMap (OSM) facilitate the labeling of many samples, the data quality is not always guaranteed, which leads to consistency issues [56]. To address this, the level of agreement among labelers is often used as a measure of consistency. For example, Fenza et al. [33] mention that simple measures, like the ratio of agreeing labelers to all labelers, can be used as an indicator of consistency. This measure can be further adjusted by considering the uncertainty and reliability of the labelers [95]. Future directions and research gaps have been identified, for example, by Saralioglu and Gungor [107].

Unbiasedness. Data creation imposes three types of bias: historical bias, representation bias, and measurement bias [14, 108]. Influencing factors for the choice of specific sampling strategies may include logistical and practical reasons for in-situ data collection. For example, as mentioned by Fowler et al. [99], in-situ crop-type labels are often collected in close proximity to roads. Even though diversity and completeness can be of high quality, biases can occur. Self-supervised learning methods like seasonal contrast [58], for example, take special care in sampling (unlabelled) images geographically according to some distribution (e.g., population density) to cover all possible representations of the world. A naive uniform sampling, though ensuring diversity to a high extent, would generate too many examples over repetitive uninformative areas like deserts or oceans. It is important to note that class imbalance in training data is a common challenge in geospatial data, which can cause the model to have a bias towards the majority class in classification tasks. This imbalance may naturally occur in some geospatial applications, where the distribution of objects or features in the studied geographic area leads to imbalanced classes. For example, in land cover classification, urban areas may be a minority class compared to forests or water bodies. Finally, measurement bias is related to how the labeling scheme is conceptualized. For example, because of the ambiguity of the class definitions, rules for labeling will always introduce bias in how objects are identified. This cannot be avoided but should be clearly communicated. Assessing and mitigating biases is a so far underrepresented research area, which however gained attention with the movement from local to global models [14].

13

Relevance. Without an existing model, there are limited works that consider a relevance measure for data creation. Instead, other quality criteria are primarily used. As a result, the collection of relevant data heavily relies on the decisions made in the problem definition step (Step 1). An approach to improve data relevance is tested in Rußwurm et al. [90], who aimed to improve urban settlement detection in Mozambique and Tanzania by dynamically sampling training data of urban areas in proximity to the study region.

To quantify the relevance of data to a given task when a model is available, one direct approach is the "leave-one-out" method [109], where the model's performance, trained on the entire dataset, is contrasted with the model's performance when a particular data point is excluded. The difference in performance can then be interpreted as the data point's relevance for the task. However, computing such values for all data points using this method can be excessively resource-intensive. A more efficient alternative to address this challenge is the use of influence functions [110, 111], which assess the influence of a specific sample on the model's parameters when that sample's weight is increased. To further address the complexity of sample interactions, the concept of Data Shapley has emerged [35], drawing inspiration from Shapley values. Data Shapley provides a more comprehensive framework for evaluating the value of individual data points in a dataset and efficient data Shapley approaches have been proposed [112]. However, such methods have not been used for geospatial data so far. In the context of incremental learning in the geospatial domain, Roscher et al. [96] uses the classification model to identify relevant samples based on their ability to help the model to adapt to the new data. More techniques that rely on an existing model are discussed in Section 3.3.

3.2. Data Curation (Step 3)

Data curation refers to the systematic and algorithmic refinement of a created dataset. We specifically focus on techniques that remove or correct data. The main objective is to improve data quality, which in turn should benefit model efficiency, accuracy, and generalizability. Similar to the data creation step, the data curation step directly impacts the quality of the data according to our defined criteria.

Diversity and completeness. A commonly used set of techniques is the filling of gaps in spatial and temporal coverage, as missing data is a common problem and causes different issues such as performance degradation,

problems in data analysis, and biased outcomes [113]. Besides classical interpolation strategies, a recent approach is the use of generative models like generative adversarial networks. Luo et al. [64], for example, impute data in one-dimensional time-series signals and outperform classical methods, and Dong et al. [65] perform inpainting in sea surface temperature images where clouds cover the scene. Another set of techniques that impute missing data is superresolution [68].

Accuracy. One of the most commonly used sets of techniques is data-cleaning, which specifically focuses on data accuracy and consistency [18]. Data-only unsupervised techniques are applied before model training and are independent of the learning objective and application. However, as discussed by Ilyas et al. and Neutatz et al. [114, 115], methods that clean the data with respect to the application goals and take the ML model into account show more promising results. Core-set selection techniques such as the one presented by Santos et al. [70] can be considered a data-cleaning approach as the selection process retains the cleanest samples. In their approach, they use clustering for satellite time series to identify samples that are mislabeled or have low accuracy with the goal of removing them from the training set to avoid a decrease in model performance. In some cases, the source of uncertainty is known and can be directly reduced. For instance, data uncertainty may arise from the presence of clouds, and many methods have been developed to automatically remove them [63, 66, 67] (see also techniques that act on diversity and completeness). Northcutt et al. [71], for example, use confident learning [103] to identify and correct label errors. Labels are flagged as noisy when their label uncertainty is high, which means the model is not confident about the label (see Section 4.3). These labeled samples are then iteratively re-labeled using a more accurate labeling function, which can be a human annotator or a more sophisticated model. When multiple labelers provide labels, similar strategies can be used to identify the most likely labels or remove highly confusing samples [26]. Most model-based approaches, however, use data with their quality information during model training instead of cleaning the data and re-training the model in two separate steps (see Model training and data utilization (Step 4)).

Consistency. Curation activities further aim to establish consistency across different datasets or data sources and annotators. Hechinger et al. [72] demonstrate that when a fixed set of labelers is given, there are often 'voting patterns' where labelers agree or disagree in a similar manner. By conducting a clustering analysis of the LCZ42 dataset [116], which was

15

labeled by multiple experts, they gained valuable insights regarding the distinguishability of classes, voting behavior, and the level of consensus based on the geographic location of the reference images. These insights can be incorporated into model training (Step 4) or used to develop curation strategies, such as removing labels with low consensus values. An open challenge is the usage of efficient labeling strategies or the identification of a subset of samples that should be annotated by multiple labelers to keep the effort low.

Unbiasedness. Furthermore, curation aims to detect and eliminate systematic biases or distortions in geospatial data. In cases where biases result from class imbalances or a non-representative spatial distribution [27], similar to resampling strategies that generate or collect new data, undersampling strategies for the majority class can be utilized. Risser et al. [73] discuss methods to remove biases in satellite images, for example, by targeted removal of specific features or samples. It is important to note that removing noisy data or data with missing values, as discussed in [113], can introduce further bias, therefore auditing strategies should be used [14] for checking. Moreover, whether class imbalance is desirable or not depends on the goals and requirements of the analysis [117].

Relevance. Curation also considers the relevance of the data. By removing irrelevant data, curation ensures that the data is tailored to specific needs and goals, focusing on the most relevant information. It also naturally reduces the dataset size, improving data efficiency. Core-set selection methods compute the distance between training and test samples and remove those with low proximity. For example, in geospatial air quality estimation, Stadtler et al. [69] demonstrate that removing the 10% of training samples with the lowest proximity to the test samples only slightly decreases test accuracy, as these samples are not relevant for training. However, they assume that the test set is representative enough for the complete data set. A more sophisticated approach is implemented and described in our first experiments (Section 4.2).

In remote sensing applications with high-dimensional data, there is often a focus on dimensionality reduction techniques. These techniques aim to reduce complexity while preserving essential information. Established feature selection techniques reduce redundancy and noise, thereby increasing model performance by focusing on the most relevant and informative features [75, 74]. With the field of explainable machine learning, new methods are introduced to calculate importance scores, sensitivities, or contributions of features and interpret them as relevance [118]. Combining these areas,

Zhang et al. [76] use a wrapper feature selection method and Shapley values to identify the most relevant multi-source remote sensing marine data features used for estimating global phytoplankton group compositions.

3.3. Model training and data utilization (Step 4)

Data set utilization is centered around optimizing the way data is used during training to improve model effectiveness. Information about the quality of the data is utilized to enhance model training. Various techniques are employed to enhance training by, e.g., speeding up convergence or improve the quality of obtained local optima. In this step, the close interplay between model-centric and data-centric learning becomes apparent, demanding for a joint development of methods for both paradignms.

Accuracy. Curriculum Learning, which performs learning from the easiest to most challenging samples, and its variants [119] have been successfully employed for several tasks in the geospatial domain. Mousavi et al. [51], for instance, introduce a novel deep curriculum learning method for the classification of PolSAR images by using an entropy-alpha target decomposition strategy to estimate the degree of complexity of each image sample. This measure of complexity is related to data accuracy since the speckle generated during the acquisition of images on vegetation targets creates a variation in the observations. Ran et al. [52] propose a curriculum learning-based strategy for unsupervised domain adaptation for a semantic segmentation task by gradually introducing more complex pseudo-labeled samples from the target domain to the training process. The complexity is given by the confidence of the samples, defined by the predicted probability. Xi et al. [54] use a combination of curriculum learning and a weighting scheme for samples in a semantic segmentation task where a domain shift arises due to distinct geographic locations of the source and target domain. Instead of only iteratively selecting the most confident pseudo-labeled samples and adding them to the training data set to adapt to the target domain, they introduce a weighting scheme that balances easy-to-transfer samples with a similar appearance as the target domain and hard classes and samples with low confidence. Dong et al. [55] propose a weighted loss function term for a landcover classification task that computes the loss with updated labels. The labels are corrected based on the uncertainty estimate for that label, which is calculated as the entropy with the largest and second-largest probability estimate of the network. Similarly, Burgert et al.[56] use self-adaptive training for multi-label remote sensing classification by adapting the targets in the loss with a moving

17

average of the given targets and the class probability estimates for them. As for the data creating step, the quantification of uncertainty and the notion of confidence is an open challenge.

Consistency. Some self-supervised learning methods utilize the information about the internal consistency within remote sensing data. Notably, Manas et al. [58] define a contrastive loss that matches samples of the same location at different times, Wang et al. use simple augmentation methods [57]. Similarly across modalities, Scheibenreif at al. [60] and Prexl and Schmitt [59] use the consistencies between optical and radar data to train a deep learning model. Also, Ayush et al. [61] show that the consistency between geolocation and image can be used to pre-train deep learning models.

Unbiasedness. Kellenberger et al. [53] employ curriculum learning to improve training strategies in a habitat suitability mapping task, where class imbalance and bias are the major challenges. In their approach, they counteract the class-imbalance by exposing the model with increasingly difficult samples, which means more imbalanced training scenarios during training.

Relevance. Yuan et al. [50] employ a self-paced curriculum learning (SPCL)-based model in combination with a weighting strategy for visual question answering (VQA) on remote sensing data. The weights are learned with an importance sampling strategy, where a large importance value means that the sample has a high relevance for the model training. Another strategy that updates the learned model with a schedule is online/incremental/continual learning. Here, the relevance and therefore the schedule is determined based on the timeliness of the samples with the goal of adapting the model to new samples that became available. In addition to timeliness, geographic relevance can also play a role. Thus, observations are not only influenced by the current imaging conditions but also by the local appearance of land cover or objects. Roscher et al. [49], for example, performs a multi-class classification of large neighboring Landsat scenes by incrementally updating the model with observations from neighboring scenes that are labeled through a self-training strategy. Bhat et al. [62] combine continual learning and curriculum learning by using the new data based on their similarity to the old classes, and additionally focussing on the accuracy of the samples from the old dataset to reduce the effect of noise and to avoid the catastrophic forgetting effect.

18

3.4. Model Evaluation (Step 5)

At this stage, the model is evaluated regarding efficiency, accuracy, and generalization ability. In general, depending on the application, a variety of metrics can be used for the model evaluation and their interpretation, including different data-specific characteristics such as class imbalance or varying prevalence, which affect the selection of the correct model [120]. We will focus on methods and tools that are used for the model evaluation with a specific focus on methods that make use of the composition of the data and data quality information.

Diversity and completeness. Following the creation of a dataset, the customary practice involves dividing samples into training, validation, and testing datasets. This division is crucial for monitoring the model's generalization and addressing common issues like overfitting or underfitting, particularly prevalent in deep learning models. While random or stratified sampling is standard in machine learning tasks, applying this directly to geospatial data can lead to spatial auto-correlation issues among nearby data points in the training and test datasets, potentially overestimating model accuracy [121, 41]. For instance, the winner of the EMCL/PKDD conference TiSeLaC land cover classification contest utilized only geographical coordinates, omitting individual spectral features in predicting land cover [122]. Thus, in geospatial applications, data sampling strategies should prioritize spatial independence. Techniques such as spatial leave-one-out cross-validation or using spatially disjoint scenes for evaluation are essential to mitigate spatial correlations and ensure robust model evaluation [40, 41]. However, Wadoux et al. [123] indicate that proper probability sampling without explicit methods to account for spatial autocorrelation is sufficient. Lyons et al. [42] compare different resampling methods for classification and show that such techniques provide robust accuracy and area estimates with confidence intervals. Techniques to identify representative sample sizes [124] for the different sets should ensure the proper evaluation of the models. In [125], the authors proposed a method for estimating the sample size required to represent the modeled or estimated dataset accurately.

A well-generalizing model should predict with a similar accuracy across the entire evaluation set. However, the model's performance often varies depending on the data distribution, which is not captured by aggregated performance measures. Simply splitting the available dataset into training, validation, and testing sets is not sufficient to ensure a reliable evaluation of the model, especially when the optimization of machine learning methods is

19

performed on a subset of data that may not represent the real data distribution, as mentioned in the previous sections. To address this issue, slicing algorithms can be used, especially for small datasets, to ensure a valuable partitioning of the data. These algorithms aim to split the data and evaluate the model on relevant sub-populations that share common characteristics. Defining these sub-populations can be challenging, but it can be based on expert domain knowledge, such as using slices with similar characteristics like geographic regions [43].

Furthermore, evaluating in and out-of-distribution is also an essential aspect of evaluating the generalization ability of the models. In-distribution evaluation is the most common way to assess the performance of the models since the predefined splits usually correspond to the same source of data, even if this set could include multiple sources and acquisition parameters. Several works discuss the usefulness and applicability of evaluation metrics [126, 127, 128]. Evaluating and challenging the performance of the models on out-of-distribution samples gain more and more attention from the community with publicly available datasets such as the WILDS [44], including also Earth observation datasets, being introduced to benchmark the performance of the algorithms on different distribution shifts. In a similar direction, the 2017 IEEE GRSS Data Fusion Contest [43] aimed at addressing the problem of local climate zone classification based on multitemporal and multimodal datasets and checking the performance of the algorithms on new and separate geographical locations. Meyer et al. [45] propose the area of applicability of a trained model by measuring and thresholding the similarity between the training data and potentially test data samples. Overall, identifying and understanding distribution shifts is a very active research area.

Accuracy. Leveraging accuracy-related information of data can be used to assess the robustness and sensitivity of machine learning models, for example, through the adoption of test-time data augmentation [28] or perturbations in the data. Such techniques introduce variations and uncertainties in the data during the evaluation phase offering an evaluation of the model's performance beyond accuracy. Li et al. [47] introduced synthetic label noise to assess the model's robustness to noise inputs. Mei et al. [46] provides a review of works that apply different adversarial perturbations to the remote sensing input images to assess the model's robustness against adversarial attacks. Kierdorf et al. [48] cluster saliency maps obtained from systematically perturbed images and the impact of these perturbations on the model decision to get insight into the reliability of the model.

4. Validation studies

The validation studies use three data-centric machine learning algorithms on a common land cover classification problem from satellite imagery: relevance-based sample weighting to compensate for geographic domain shift, label noise reduction and confident learning, and slicing for a comprehensive evaluation. Each of them highlight a different aspect of data-centric learning and provides both qualitative images highlighting the effect of the procedure and quantitative results that the respective algorithm has on the land cover classification problem. They confirm the functionality of existing approaches within the surveyed field and should emphasize that even single data-centric actions in the machine learning pipeline can lead to better models.

4.1. DFC2020 Land Cover Classification and Experimental Setup

We chose land cover classification to illustrate different data-centric approaches and use the IEEE GRSS Data Fusion Contest 2020 (DFC2020) dataset [129]. The DFC2020 addressed large-scale land cover mapping with weak supervision. The training data is Sen12MS [7], which contains only very low-resolution MODIS-derived annotations. However, for the additional validation data (i.e., not contained in the Sen12MS dataset), semi-manually generated high-resolution labels are provided. We use the Sentinel-2 images from this part of the DFC2020 dataset for the following experiments. The Sentinel-2 images are provided with 10 bands (where the native resolution of 10m and 20m is resampled to 10m) and are acquired over seven globally distributed regions. They were semi-manually annotated at high-resolution (10m) according to a simplified International Geosphere-Biosphere Programme (IGBP) classification scheme with eight classes "Forest", "Shrubland", "Grassland", "Wetlands", "Croplands", "Urban/Built-up", "Barren", and "Water". While DFC2020 is originally a segmentation dataset, we simplify it for classification by taking the most common labelled pixel as the single label for each image.

This dataset offers interesting characteristics that enable us to showcase different aspects of the potential of data-centric learning. On the one hand, the image data is spatially well distributed over different climate zones, biomes, and socio-cultural regions. This results in data characteristics that will be shared only among a subset of images. A similar diversity can also be expected from the semantic content of the data as captured by the annotations. On the other hand, parts of the data will be easier to correctly

classify than others as they are better represented by prototypical samples and are more homogeneous in their appearance (e.g. samples of the "Barren" or "Water" class). Furthermore, while the annotations are created in a semi-automatic process, i.e. with thorough quality control, they still contain a certain amount of noise, ambiguity, and errors.

The seven scenes are split into non-overlapping 256×256 pixel patches, resulting in a total of 6,114 images. As we are dealing with classification tasks for the following experiments, we follow the classification-oriented conversion proposed by the creator of dataset and determine the dominant label (i.e., the class with the most pixels) within a patch and use it as its patch-level label [130]. These samples are partitioned randomly into 4,000 images for training, 1,128 for validation, and 986 images for testing in the Experiments 2 and 3. In Experiment 1, we split the images by region to specifically mitigate and correct for the distribution shift between regions. In this first experiment, we also use a common Random Forest classification model, while for Experiments 2-3, we use a deep classification baseline model and train a ResNet-18 convolutional neural network with the AdamW optimizer using a learning rate of 0.001 and weight decay of 10^{-5} with up to 1000 epochs. We saved the model with the lowest validation loss, which resulted in a model with 123 epochs after four hours of training.

4.2. Study 1: Relevance-based sample weighting to compensate for geographic domain shift

Motivation. Varying environmental and geographic conditions changes the visual representation of land cover classes. This can be seen in Fig. 3a where the average reflectance in red and near-infrared channels varies systematically between the DFC2020 regions indicated by color. This manifests a data distribution shift between geographic regions, negatively impacting model classification accuracy when not all training data is equally *relevant* for the particular testing region. In this case, data from the source distribution (i.e., training data) does not overlap with the target distribution (testing data). This can be seen in the case of Cape Town and Black Forest in Fig. 3a.

Aim. We aim to reduce the distribution shift with respect to a target region by weighting each image sample with an importance coefficient in an instance-based transfer learning approach [131]. In this method, samples from the source distribution that are assigned high importance weights are deemed more *relevant* for the particular target distribution, while low weights are assigned to irrelevant samples.
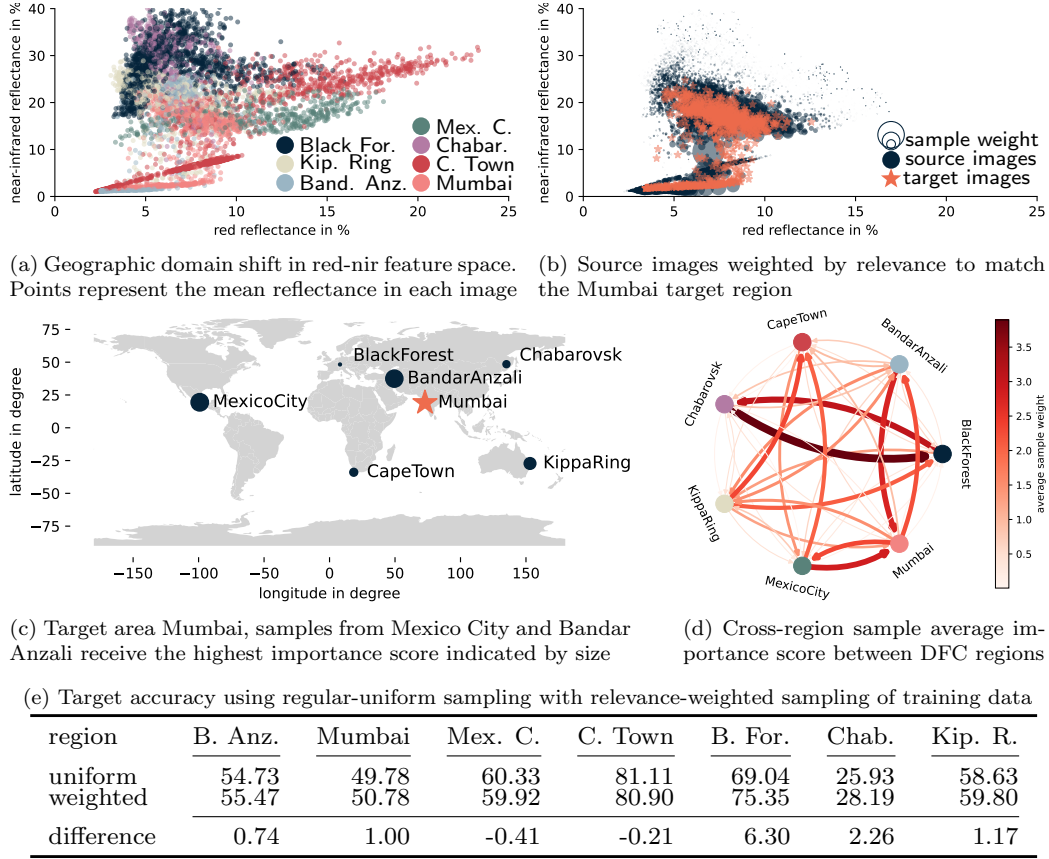
(a) Geographic domain shift in red-nir feature space. Points represent the mean reflectance in each image

(b) Source images weighted by relevance to match the Mumbai target region

(c) Target area Mumbai, samples from Mexico City and Bandar Anzali receive the highest importance score indicated by size

(d) Cross-region sample average importance score between DFC regions

(e) Target accuracy using regular-uniform sampling with relevance-weighted sampling of training data

region	B. Anz.	Mumbai	Mex. C.	C. Town	B. For.	Chab.	Kip. R.
uniform	54.73	49.78	60.33	81.11	69.04	25.93	58.63
weighted	55.47	50.78	59.92	80.90	75.35	28.19	59.80
difference	0.74	1.00	-0.41	-0.21	6.30	2.26	1.17

Figure 3: Experiment 1: Sample-wise debiasing geographic domain shift with KLIEP.

Method. We use the Kullback–Leibler importance estimation procedure (KLIEP) [132] using the Frank-Wolfe algorithm [133] implemented in the Python Adapt package [134] to determine the relevance-weighting scheme. This algorithm approximates the source and target distributions by the sum of each sample's radial basis function (RBF) kernel densities. We perform a grid-search for kernel width between 0.1 and 2.0 in regular 0.1 steps. KLIEP iteratively adjusts the sample weights so that Kullback-Leibler (KL) divergence between target distribution and source distribution is minimized, as shown in Fig. 3b. We represent each image sample as a 20-dimensional vector of pixel-average band reflectances (10 channels) and their standard deviations (10 channels). We then fit a random forest classifier from Scikit-Learn [135] with uniform weighting and a KLIEP-derived importance weighting on the

Table 1: Target accuracy in [%] using regular-uniform weighting vs. relevance-weighting of training data

Region	B. Anz.	Mumbai	Mex. C.	C. Town	B. For.	Chab.	Kip. R.
Uniform sampling	54.73	49.78	60.33	81.11	69.04	25.93	58.63
Weighted sampling	55.47	50.78	59.92	80.90	75.35	28.19	59.80
Difference	0.74	1.00	-0.41	-0.21	6.30	2.26	1.17

source regions and measure the accuracy on the target region.

Results and Interpretation. We show results in Figs. 3c and 3d and Table 1. First, Fig. 3c shows a world map with average relevance of all images within each source DFC region (scaled circles) with respect to Mumbai. This generally covers Tobler's [136] intuition that more geographically nearby areas, like Bandar Anzali, are more relevant than distant ones. However, environmental conditions beyond mere geographic distance also contribute to the relevance, as the high importance of Mexico City images shows that it shares a similar arid climate. Second, Fig. 3d extends this comparison across all regions and shows the sample-averaged importance scores as edge width and color in a bi-directional graph between all DFC different regions. Here, we can see more generally that nearby (Mumbai ↔ Bandar Anzali) temperate and forested regions (Black Forest ↔ Chabaraovsk) and arid regions (Cape Town ↔ Mexico City ↔ Mumbai ↔ Kippa Ring) are more relevant for each other.

Finally, Table 1 quantifies the accuracy benefit of classifying on a sample-weighted training source dataset with respect to each target region in the columns. The row "uniform weighting" shows classification results on the random forest trained on regular uniformly weighted training data, while "relevance weighting" reports the accuracy on KLIEP-weighted training samples. The "difference" rows indicate in which regions the de-biasing was beneficial (positive) or detrimental (negative) to the final target accuracy. Here, the weighting positively affected five of the seven DFC regions up to 6.3%. Two regions were classified at lower accuracy (up to -0.41%). That this procedure can also have detrimental effects in same case illustrates the inherent challenge of deriving the relevance of samples and that negative transfer can occur and deteriorate performance.

Summary. We implement and demonstrate the KLIEP algorithm to derive a relevance-based weighting scheme to account for geographic distribution shifts. This reveals patterns in the data, such as which geographic areas
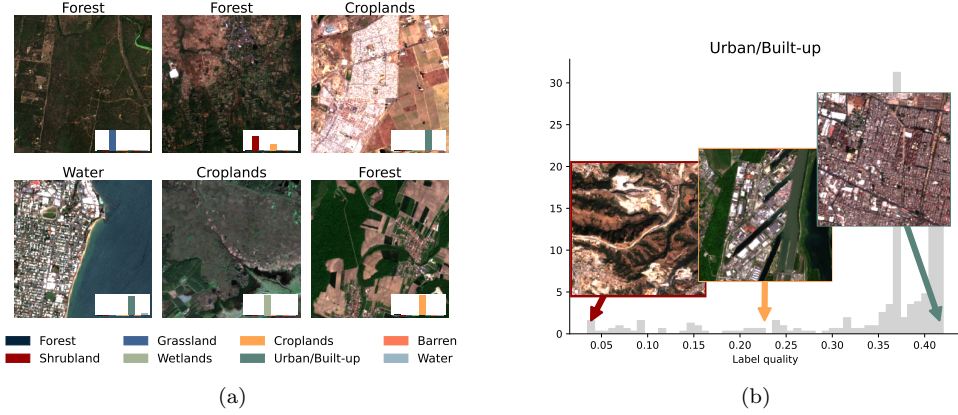
Figure 4: Qualitative results, Experiment 2. (a) Potential label issues detected by Confident Learning in the DFC training set. The model's predicted class probabilities are shown as bar charts. (b) Histogram of the label quality of class "Urban/Built-up" with three example images.

are more *classification-relevant* than others, but also improves the predictive accuracy when weighting the training samples accordingly. However, this procedure does not always increase the accuracy. It can also have detrimental effects when less relevant samples are falsely assigned a higher importance/relevance score due to incomplete approximations and simplifications, as is necessary for choosing kernel functions or the distribution divergence metric with computational restrictions in mind. Further research is necessary to make these algorithms applicable with less supervision and more general with less hyperparameter tuning.

4.3. Study 2: Pruning noisy data with confident learning

Table 2: Quantitative results of study 2. Model performance when training on the entire training dataset (baseline) and after pruning noisy training samples (confident learning).

	Accuracy [%]
Baseline	66.53
Confident learning	69.98

Motivation. *Accurate* data labels are crucial for the effectiveness of trained models, yet inaccuracies can significantly undermine model performance. In geospatial remote sensing, label issues often arise due to similarities in appearance between classes, such as "Grassland" and "Shrubland" sharing com-

25

parable vegetation characteristics, which can lead to confusion. Additionally, the spatial distribution of certain classes, like "Wetlands" being commonly located near water, increases the likelihood of co-occurrence within patches, further complicating accurate labeling.

Aim. Our aim is to improve label accuracy by ranking data samples and pruning potentially noisy samples

Method. To achieve this, we employ confident learning [71], a framework designed to estimate label uncertainty and enhance dataset quality. Confident learning operates under the assumption that label noise is class-conditional, depending solely on the true class rather than the data itself [103]. Although this is a simplification, it is a reasonable assumption for many scenarios, including land cover classification.

The method involves training a machine learning model on the dataset and then estimating the joint distribution of noisy (given) labels and true (unknown) labels. This estimation is done by identifying examples likely to have label issues. For instance, it counts all patches labeled as "Grassland" that have a sufficiently high predicted probability of being "Shrubland". Here, a value for sufficiently high is determined by a class-specific threshold calculated as the model's average self-confidence for each class. Using the estimated joint distribution, all samples are ranked based on their likelihood of being noisy. Potentially noisy samples are then pruned from the dataset.

Results and Interpretation. The confident learning method flagged 123 samples in the DFC training set as having potential label issues. Six of these samples are presented in Fig. 4a, illustrating instances where the model disagreed with the given label. In some cases, as seen in the lower left image, discrepancies arise from the presence of multiple classes within one patch, while in other cases, such as the top left image, the reason is related to the similarity in appearance between classes. Figure 4b shows a histogram of the label quality of all training samples belonging to class "Urban/Built-up", accompanied by three example images representing different label qualities: one with high label quality, one with medium label quality, and one with low label-quality. These examples show that the label quality tends to degrade as the fraction of "Urban/Built-up" pixels within the patch decreases. To improve the quality of the training data, the samples flagged as noisy were removed. Re-training the model on this pruned dataset resulted in an overall accuracy improvement of +3.45 percentage points compared to the baseline (see Table 2).

Summary. Confident learning is a form of data curation that is typically used to improve the training dataset quality by pruning noisy data. This approach is especially relevant for settings with low label granularity, where multiple classes are present within the same image. By pruning the dataset, confident learning aims to improve the *accuracy* of the labels.

4.4. Study 3: Evaluation with slice discovery

Motivation. To ensure a comprehensive evaluation of model performance, it is beneficial to identify human-understandable subsets of test samples within each class, where the model's performance varies across different slices. This approach provides deeper insights into the model's strengths and weaknesses.
Aim. Our aim is to demonstrate the insights that can be gained by analyzing slices in the test data.
Method. We utilize the idea of slice-finding methods, such as presented in [137]. To analyze the data at test time, we first represent all data points by a latent representation vector, produced via feature extraction of a pre-trained *ResNet18* neural network of Prexl et al. [59], which is a domain-specific (trained on *Sentinel-2* images of the *SEN12MS* dataset) pre-trained network based on the SimCLR contrastive learning paradigm [138]. We extract a 512 dimensional feature for each image in the test set of the *DFC2020* dataset. Using those latent feature representations allows for sub-clustering of all samples within one class by applying *k-means* clustering in the latent space.
Results and Interpretation. The cluster-center affiliation of each sample allows for quick and robust visual analysis of the data and can produce further insights about the task at hand. This can subsequently be used to make informed decisions regarding improved labeling schemes or feedback into the data curation step (Step 2 in Fig. 1). We show two examples of "Cropland" and "Water" with their dominant sub-slices in Fig. 6. This reveals directly different patterns within one class label, which might harm the model's performance. In the first case, looking at all samples labeled as "Water", it can be observed that the slices capture different subcategories like river, lake, and open ocean. While "Water" is accurately classified, "Cropland" is commonly confused with "Grassland", as shown in the confusion matrix Fig. 5. Here, analyzing sub-slices can reveal failure cases: "Cropland" samples cluster in two slices with different climate conditions or seasons. While the performance on all samples in the first slice (Fig. 6) has 100% accuracy, the performance
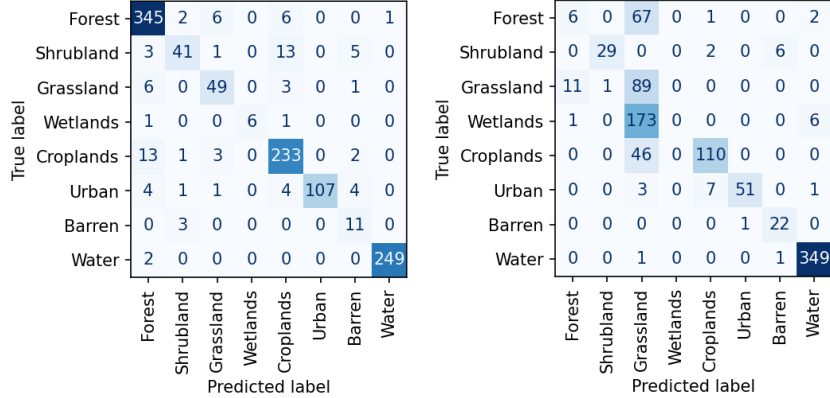
Figure 5: The confusion matrices (left: validation set, right: test set) of the model described in Section 4.1 for validation (left) and test set (right).

for the second slice drops significantly to 10%, likely due to the high similarity to the "Grassland" samples it he training set.

Summary. Overall, slice discovery methods serve as tools to gain insights into the outcome of a specific task that is given to the model beyond accumulated numbers. This insight can be used directly for data creation and curation to improve the performance by, for example, creating new classes for the sub-slices.

5. Discussion & Conclusion

Shifting research focus from well-explored model-centric approaches towards the broader aspects of the entire machine learning cycle, especially data-centric learning techniques, is necessary to make further progress towards accurate and generalizable models. Here, we find that so far, only limited research is focused on automated data-centric approaches that explicitly improve the quality criteria that we categorized in Section 2 or utilize them for the model-building process. In this work, we provided an extensive literature review in Section 3 and categorized the techniques according to steps in the machine learning cycle and identified the quality criteria they improve or use. This literature study was complemented by concrete implementation examples in Section 4 to stress the applicability of some of the data-centric methods.

Research Gaps. Unlike the numerous well-established metrics available to
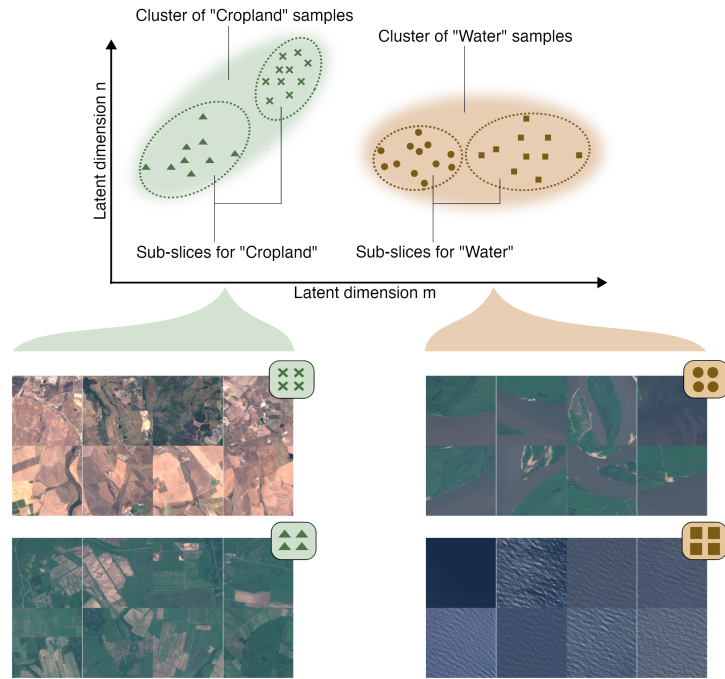
Figure 6: The cluster analysis showcases representative slices of all samples for the "Cropland" and "Water" classes. By clustering the latent representations, distinct sub-clusters of cropland and water types are identified within the test set of the dataset. The top row presents a schematic example of the clustering results, while the bottom two rows offer visual examples of the sub-cluster analysis for the "Cropland" and "Water" classes.

measure the performance of machine learning models, standardized metrics to evaluate the quality of a dataset are not established so far [1]. Within the studied data-centric approaches, we identified research gaps in Fig. 2 in the following steps:

Step 2: Creation Only a few existing works focus on automated data creation to improve data consistency, unbiasedness, or relevance. While these criteria are often considered in the manual selection of training and evaluation sites, little work has been found on the automation of this process through data-centric learning in the geospatial domain. Also, little research has been conducted so far on the benefit of synthetic, generated, and augmented data for geospatial remote sensing data.

Step 3: Curation We find that few works focus on maximizing diversity, consistency and/or unbiasedness through data curation. This has several factors: First, it is plausible for diversity, as curation is more aimed at cleaning up and deleting data, which hardly increases diversity. Moreover, it is challenging to quantify diversity and completeness as a reference is missing. Usually, such shortcomings are revealed after the model deployment step, which can be fed back to the machine learning cycle. Currently, however, most works do not consider feedback loops. Research related to consistency is restricted due to the limited availability of datasets (i.e., multi-labeler datasets) that allow focusing on this aspect. Bias is still an under-investigated criterium and its cascading effects on the results are underestimated. Usually, biases in the dataset are revealed too late in the model deployment (Step 6). Fortunately, a new sensitivity has evolved towards this criteria. Overall, research is currently hindered due to the lack of clear objectives on what a data set should look like regarding the quality criteria. For example, curation strategies can be too time consuming and costly compared to their benefit on the dataset. Also data-centric and model-centric developments need to be balanced to find the best way to deal with the data and the application.

Step 4: Utilization We find that few works use completeness or unbiasedness. However, in contrast to previous steps, we believe that these criteria are less relevant for the data utilization in the model training

context, while diversity, accuracy, and relevance are well covered by existing literature.

Step 5: Evaluation As already underlined by Rolf et al. [128], there are several opportunities for a better evaluation that have not been exploited yet. Investing in a high-quality evaluation data set is important to get a comprehensive view on the ability of the model. Further, we need to distinguish between performance measures reported with a statistical parameter and measures that provide insights into the skill for a given task. Generally, there is limited research that uses consistency, bias, and relevance information of the data to get insights into the model's performance.

Experimental Performance. Another important point of discussion is the effectiveness of current data-centric approaches in experiments. While all proposed methods are well-motivated, the actual improvement in accuracy often falls short of expectations. For example, in Study 1 (Section 4.2), weighting dataset samples regarding their relevance only improves two out of five regions. Similarly, in Study 2 (Section 4.3), the quantitative accuracy gain by removing samples with label noise only results in a 3.5% improvement. However, both approaches reveal underlying patterns in the data that can provide valuable insights on their own. Additionally, in real-world applications, accuracy is not the sole factor of interest. Factors such as reliability, precision, robustness, and adaptability are often equally important. Furthermore, a major goal is to avoid the negative effects caused by low-quality data [17]. We further want to point out that the used methods were chosen due to their already proven performance in diverse fields. However, the success of data-centric techniques also depend on the chosen model, therefore we encourage considering the whole machine learning cycle, including model- and data-centric aspects.

Limitations. Several limitations need to be considered in the interpretation of this study. First, the proposed categorizations are not sharp. Methods can be argued for improving multiple criteria. For instance, the first study uses KLIEP [132] (data curation) to address the relevance, but it can also be seen as a de-biasing strategy. Here, the irrelevant samples with respect to one target area receive a low weight and are discarded. Simultaneously, this process of selecting the most relevant instances de-biases the entire datasets

31

if we are aware of a certain sampling bias, e.g., towards some continents. Similarly, the process of data augmentation can be argued towards data creation (step 2), as new augmented samples are virtually created, or data utilization (step 4), as the existing data is utilized more effectively by flipping and rotating images without the need to collect new data samples. Finally, we highlight that the manual search of literature always introduces a certain selection bias. We aimed to mitigate this by inviting a diverse range of co-authors who work across disciplines that create, curate, and utilize geospatial remote sensing data. We believe that this broad expertise allowed us to extensively and intensively search the entire body of published literature in this domain.

Conclusion and Outlook. We consider this review as a starting point to raise awareness about the possibilities of data-centric machine learning in the geospatial domain rather then arguing against model-centric research. This review introduced the definitions, and relevant criteria to give researchers the tools and structure to improve datasets in an automated and systematic way. Focusing on the entire machine learning cycle from problem definition to model deployment with feedback is necessary to train machine learning models that can generalize well to new locations and can be trusted in unexpected situations. This is captured and quantified in data-centric learning that explicitly improves and uses the relevant data quality criteria. Future research is necessary to fill the identified research gaps in data creation and curation towards a higher quality of datasets. Further, additional methodological developments are necessary to improve the reliability of data-centric methods that today often fall behind expectations. In summary, we hope that this review helps to structure and guide this discussion and leads to method developments targeted towards new unexplored directions that help tackle real-world problems with high-quality data.

6. Acknowledgments

References

[1] L. Aroyo, M. Lease, P. Paritosh, and M. Schaekermann, "Data excellence for ai: why should you care?," *Interactions*, vol. 29, no. 2, pp. 66–69, 2022.

[2] N. Polyzotis and M. Zaharia, "What can data-centric ai learn from data and ml engineering?," in *Proc. of the Conference on Neural Information Processing Systems*, 2021.

[3] K. L. Wagstaff, "Machine learning that matters," in *Proc. of the International Coference on International Conference on Machine Learning*, pp. 1851–1856, 2012.

[4] L. Zhang and L. Zhang, "Artificial intelligence for remote sensing data analysis: A review of challenges and opportunities," *IEEE Geoscience and Remote Sensing Magazine*, vol. 10, no. 2, pp. 270–294, 2022.

[5] I. Kraljevski, C. Tschöpe, and M. Wolff, "Limits and prospects of big data and small data approaches in ai applications," *KI-Kritik/AI Critique Volume 4*, p. 115, 2023.

[6] M. Schmitt, S. A. Ahmadi, Y. Xu, G. Taşkın, U. Verma, F. Sica, and R. Hänsch, "There are no data like more data: Datasets for deep learning in earth observation," *IEEE Geoscience and Remote Sensing Magazine*, 2023.

[7] 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," in *ISPRS Ann. Photogramm. Remote Sens. Spat. Inf. Sci.*, vol. IV-2/W7, pp. 153–160, 2019.

[8] V. S. F. Garnot and L. Landrieu, "Panoptic segmentation of satellite image time series with convolutional temporal attention networks," in

Proc. of the IEEE/CVF International Conference on Computer Vision, pp. 4872–4881, 2021.

[9] M. Goldblum, A. Anandkumar, R. Baraniuk, T. Goldstein, K. Cho, Z. C. Lipton, M. Mitchell, P. Nakkiran, M. Welling, and A. G. Wilson, "Perspectives on the state and future of deep learning–2023," *arXiv preprint arXiv:2312.09323*, 2023.

[10] D. Hong, B. Zhang, X. Li, Y. Li, C. Li, J. Yao, N. Yokoya, H. Li, P. Ghamisi, X. Jia, *et al.*, "Spectralgpt: Spectral remote sensing foundation model," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[11] D. Wang, M. Hu, Y. Jin, Y. Miao, J. Yang, Y. Xu, X. Qin, J. Ma, L. Sun, C. Li, *et al.*, "Hypersigma: Hyperspectral intelligence comprehension foundation model," *arXiv preprint arXiv:2406.11519*, 2024.

[12] G. Mai, W. Huang, J. Sun, S. Song, D. Mishra, N. Liu, S. Gao, T. Liu, G. Cong, Y. Hu, *et al.*, "On the opportunities and challenges of foundation models for geospatial artificial intelligence," *arXiv preprint arXiv:2304.06798*, 2023.

[13] X. Sun, P. Wang, W. Lu, Z. Zhu, X. Lu, Q. He, J. Li, X. Rong, Z. Yang, H. Chang, *et al.*, "Ringmo: A remote sensing foundation model with masked image modeling," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–22, 2022.

[14] P. Ghamisi, W. Yu, A. Marinoni, C. M. Gevaert, C. Persello, S. Selvakumaran, M. Girotto, B. P. Horton, P. Rufin, P. Hostert, *et al.*, "Responsible ai for earth observation," *arXiv preprint arXiv:2405.20868*, 2024.

[15] A. Ng, "Unbiggen ai," *IEEE Spectrum*, vol. 9, 2022.

[16] M. H. Jarrahi, A. Memariani, and S. Guha, "The principles of data-centric ai," *Communications of the ACM*, vol. 66, no. 8, pp. 84–92, 2023.

[17] N. Sambasivan, S. Kapania, H. Highfill, D. Akrong, P. Paritosh, and L. M. Aroyo, ""everyone wants to do the model work, not the data

work": Data cascades in high-stakes ai," in *Proc. of the CHI Conference on Human Factors in Computing Systems*, pp. 1–15, 2021.

[18] P. Li, X. Rao, J. Blase, Y. Zhang, X. Chu, and C. Zhang, "Cleanml: A study for evaluating the impact of data cleaning on ml classification tasks," in *Proc. of the IEEE International Conference on Data Engineering (ICDE)*, pp. 13–24, 2021.

[19] S. Falk and A. van Wynsberghe, "Challenging ai for sustainability: what ought it mean?," *AI and Ethics*, pp. 1–11, 2023.

[20] R. Hadsell, D. Rao, A. A. Rusu, and R. Pascanu, "Embracing change: Continual learning in deep neural networks," *Trends in Cognitive Sciences*, vol. 24, no. 12, pp. 1028–1040, 2020.

[21] A. K. Aksoy, M. Ravanbakhsh, and B. Demir, "Multi-label noise robust collaborative learning for remote sensing image classification," *IEEE Transactions on Neural Networks and Learning Systems*, 2022.

[22] K. Fatras, B. B. Damodaran, S. Lobry, R. Flamary, D. Tuia, and N. Courty, "Wasserstein adversarial regularization for learning with label noise," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 10, pp. 7296–7306, 2021.

[23] L. Zhang, Y. Zhao, Z. Zhu, D. Shen, and S. Ji, "Multi-view missing data completion," *IEEE Transactions on Knowledge and Data Engineering*, vol. 30, no. 7, pp. 1296–1309, 2018.

[24] Y. Chen, R. Cao, J. Chen, L. Liu, and B. Matsushita, "A practical approach to reconstruct high-quality landsat ndvi time-series data by gap filling and the savitzky–golay filter," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 180, pp. 174–190, 2021.

[25] G. Machado, M. B. Pereira, K. Nogueira, and J. A. dos Santos, "Facing the void: Overcoming missing data in multi-view imagery," *IEEE Access*, vol. 11, pp. 12547–12554, 2023.

[26] Y. Long, G.-S. Xia, S. Li, W. Yang, M. Y. Yang, X. X. Zhu, L. Zhang, and D. Li, "On creating benchmark dataset for aerial image interpretation: Reviews, guidances, and million-aid," *IEEE Journal of Selected*

Topics in Applied Earth Observations and Remote Sensing, vol. 14, pp. 4205–4230, 2021.

[27] S. Shankar, Y. Halpern, E. Breck, J. Atwood, J. Wilson, and D. Sculley, "No classification without representation: Assessing geodiversity issues in open data sets for the developing world," in *Proc. of the NeurIPS Workshop on Machine Learning for the Developing World*, 2017.

[28] J. Gawlikowski, C. R. N. Tassi, M. Ali, J. Lee, M. Humt, J. Feng, A. Kruspe, R. Triebel, P. Jung, R. Roscher, *et al.*, "A survey of uncertainty in deep neural networks," *Artificial Intelligence Review*, vol. 56, pp. 1513–1589, 2023.

[29] A. Elmes, H. Alemohammad, R. Avery, K. Caylor, J. R. Eastman, L. Fishgold, M. A. Friedl, M. Jain, D. Kohli, J. C. Laso Bayas, *et al.*, "Accounting for training data error in machine learning applied to earth observations," *Remote Sensing*, vol. 12, no. 6, p. 1034, 2020.

[30] C. M. Gevaert, D. Kohli, and M. Kuffer, "Challenges of mapping the missing spaces," in *Proc. of the Joint Urban Remote Sensing Event (JURSE)*, pp. 1–4, 2019.

[31] B. Ekim, T. T. Stomberg, R. Roscher, and M. Schmitt, "Mapinwild: A remote sensing dataset to address the question of what makes nature wild [software and data sets]," *IEEE Geoscience and Remote Sensing Magazine*, vol. 11, no. 1, pp. 103–114, 2023.

[32] T. T. Stomberg, J. Leonhardt, I. Weber, and R. Roscher, "Recognizing protected and anthropogenic patterns in landscapes using interpretable machine learning and satellite imagery," *Frontiers in Artificial Intelligence*, vol. 6, 2023.

[33] G. Fenza, M. Gallo, V. Loia, F. Orciuoli, and E. Herrera-Viedma, "Data set quality in machine learning: consistency measure based on group decision making," *Applied Soft Computing*, vol. 106, p. 107366, 2021.

[34] J. E. Vargas-Munoz, S. Srivastava, D. Tuia, and A. X. Falcao, "Openstreetmap: Challenges and opportunities in machine learning and remote sensing," *IEEE Geoscience and Remote Sensing Magazine*, vol. 9, no. 1, pp. 184–199, 2020.

[35] A. Ghorbani and J. Zou, "Data shapley: Equitable valuation of data for machine learning," in *Proc. of the International Conference on Machine Learning*, pp. 2242–2251, PMLR, 2019.

[36] J. T. Wang and R. Jia, "Data banzhaf: A robust data valuation framework for machine learning," in *Proc. of the International Conference on Artificial Intelligence and Statistics*, pp. 6388–6421, PMLR, 2023.

[37] X. Xu, Z. Wu, C. S. Foo, and B. K. H. Low, "Validation free and replication robust volume-based data valuation," *Advances in Neural Information Processing Systems*, vol. 34, pp. 10837–10848, 2021.

[38] F. Schomm, F. Stahl, and G. Vossen, "Marketplaces for data: an initial survey," *ACM SIGMOD Record*, vol. 42, no. 1, pp. 15–26, 2013.

[39] D. Tuia, M. Volpi, L. Copa, M. Kanevski, and J. Munoz-Mari, "A survey of active learning algorithms for supervised remote sensing image classification," *IEEE Journal of Selected Topics in Signal Processing*, vol. 5, no. 3, pp. 606–617, 2011.

[40] R. Hänsch, "The trap of random sampling and how to avoid it-alternative sampling strategies for a realistic estimate of the generalization error in remote sensing," in *Proc. of the IEEE International Geoscience and Remote Sensing Symposium*, pp. 2020–2023, 2021.

[41] N. Karasiak, J.-F. Dejoux, C. Monteil, and D. Sheeren, "Spatial dependence between training and test sets: another pitfall of classification accuracy assessment in remote sensing," *Machine Learning*, vol. 111, no. 7, pp. 2715–2740, 2022.

[42] M. B. Lyons, D. A. Keith, S. R. Phinn, T. J. Mason, and J. Elith, "A comparison of resampling methods for remote sensing classification and accuracy assessment," *Remote Sensing of Environment*, vol. 208, pp. 145–153, 2018.

[43] N. Yokoya, P. Ghamisi, J. Xia, S. Sukhanov, R. Heremans, I. Tankoyeu, B. Bechtel, B. Le Saux, G. Moser, and D. Tuia, "Open data for global multimodal land use classification: Outcome of the 2017 ieee grss data fusion contest," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 11, no. 5, pp. 1363–1377, 2018.

[44] P. W. Koh, S. Sagawa, H. Marklund, S. M. Xie, M. Zhang, A. Balsub-ramani, W. Hu, M. Yasunaga, R. L. Phillips, I. Gao, *et al.*, "Wilds: A benchmark of in-the-wild distribution shifts," in *Proc. of the International Conference on Machine Learning*, pp. 5637–5664, PMLR, 2021.

[45] H. Meyer and E. Pebesma, "Predicting into unknown space? estimating the area of applicability of spatial prediction models," *Methods in Ecology and Evolution*, vol. 12, no. 9, pp. 1620–1633, 2021.

[46] S. Mei, J. Lian, X. Wang, Y. Su, M. Ma, and L.-P. Chau, "A comprehensive study on the robustness of image classification and object detection in remote sensing: Surveying and benchmarking," *arXiv preprint arXiv:2306.12111*, 2023.

[47] Y. Li, Y. Zhang, and Z. Zhu, "Error-tolerant deep learning for remote sensing image scene classification," *IEEE Transactions on Cybernetics*, vol. 51, no. 4, pp. 1756–1768, 2020.

[48] J. Kierdorf and R. Roscher, "Reliability scores from saliency map clusters for improved image-based harvest-readiness prediction in cauliflower," *IEEE Geoscience and Remote Sensing Letters*, vol. 20, pp. 1–5, 2023.

[49] R. Roscher, B. Waske, and W. Förstner, "Incremental import vector machines for large area land cover classification," in *Proc. of the IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*, pp. 243–248, 2011.

[50] Z. Yuan, L. Mou, Q. Wang, and X. X. Zhu, "From easy to hard: Learning language-guided curriculum for visual question answering on remote sensing data," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–11, 2022.

[51] H. Mousavi, M. Imani, and H. Ghassemian, "Deep curriculum learning for polsar image classification," in *Proc. of the IEEE International Conference on Machine Vision and Image Processing (MVIP)*, pp. 1–5, 2022.

[52] L. Ran, C. Ji, S. Zhang, X. Zhang, and Y. Zhang, "An unsupervised domain adaption framework for aerial image semantic segmentation based

on curriculum learning," in *Proc. of the IEEE International Conference on Image, Vision and Computing (ICIVC)*, pp. 354–359, 2022.

[53] B. Kellenberger, E. Cole, D. Marcos, and D. Tuia, "Training techniques for presence-only habitat suitability mapping with deep learning," in *Proc. of the IEEE International Geoscience and Remote Sensing Symposium*, pp. 5085–5088, IEEE, 2022.

[54] Z. Xi, X. He, Y. Meng, A. Yue, J. Chen, Y. Deng, and J. Chen, "A multilevel-guided curriculum domain adaptation approach to semantic segmentation for high-resolution remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, 2023.

[55] R. Dong, W. Fang, H. Fu, L. Gan, J. Wang, and P. Gong, "High-resolution land cover mapping through learning with noise correction," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–13, 2021.

[56] T. Burgert, M. Ravanbakhsh, and B. Demir, "On the effects of different types of label noise in multi-label remote sensing image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–13, 2022.

[57] Y. Wang, N. A. A. Braham, Z. Xiong, C. Liu, C. M. Albrecht, and X. X. Zhu, "Ssl4eo-s12: A large-scale multi-modal, multi-temporal dataset for self-supervised learning in earth observation," *arXiv preprint arXiv:2211.07044*, 2022.

[58] O. Manas, A. Lacoste, X. Giró-i Nieto, D. Vazquez, and P. Rodriguez, "Seasonal contrast: Unsupervised pre-training from uncurated remote sensing data," in *Proc. of the IEEE/CVF International Conference on Computer Vision*, pp. 9414–9423, 2021.

[59] J. Prexl and M. Schmitt, "Multi-modal multi-objective contrastive learning for sentinel-1/2 imagery," in *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 2136–2144, 2023.

[60] L. Scheibenreif, J. Hanna, M. Mommert, and D. Borth, "Self-supervised vision transformers for land-cover segmentation and classi-

fication," in *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1421–1430, 2022.

[61] K. Ayush, B. Uzkent, C. Meng, K. Tanmay, M. Burke, D. Lobell, and S. Ermon, "Geography-aware self-supervised learning," in *Proc. of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10181–10190, October 2021.

[62] S. D. Bhat, B. Banerjee, S. Chaudhuri, and A. Bhattacharya, "Efficient curriculum based continual learning with informative subset selection for remote sensing scene classification," *arXiv preprint arXiv:2309.01050*, 2023.

[63] Y. Zhang, F. Wen, Z. Gao, and X. Ling, "A coarse-to-fine framework for cloud removal in remote sensing image sequence," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 8, pp. 5963–5974, 2019.

[64] Y. Luo, X. Cai, Y. Zhang, J. Xu, *et al.*, "Multivariate time series imputation with generative adversarial networks," *Advances in Neural Information Processing Systems*, vol. 31, 2018.

[65] J. Dong, R. Yin, X. Sun, Q. Li, Y. Yang, and X. Qin, "Inpainting of remote sensing sst images with deep convolutional generative adversarial network," *IEEE Geoscience and Remote Sensing Letters*, vol. 16, no. 2, pp. 173–177, 2018.

[66] J. Li, Z. Wu, Z. Hu, J. Zhang, M. Li, L. Mo, and M. Molinier, "Thin cloud removal in optical remote sensing images based on generative adversarial networks and physical model of cloud distortion," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 166, pp. 373–389, 2020.

[67] P. Ebel, Y. Xu, M. Schmitt, and X. X. Zhu, "Sen12ms-cr-ts: A remote-sensing data set for multimodal multitemporal cloud removal," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–14, 2022.

[68] P. Wang, B. Bayram, and E. Sertel, "A comprehensive review on deep learning based remote sensing image super-resolution methods," *Earth-Science Reviews*, p. 104110, 2022.

[69] S. Stadtler, C. Betancourt, and R. Roscher, "Explainable machine learning reveals capabilities, redundancy, and limitations of a geospatial air quality benchmark dataset," *Machine Learning and Knowledge Extraction*, vol. 4, no. 1, pp. 150–171, 2022.

[70] L. A. Santos, K. R. Ferreira, G. Camara, M. C. Picoli, and R. E. Simoes, "Quality control and class noise reduction of satellite image time series," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 177, pp. 75–88, 2021.

[71] C. Northcutt, L. Jiang, and I. Chuang, "Confident learning: Estimating uncertainty in dataset labels," *Journal of Artificial Intelligence Research*, vol. 70, pp. 1373–1411, 2021.

[72] K. Hechinger, X. X. Zhu, and G. Kauermann, "Categorising the world into local climate zones–towards quantifying labelling uncertainty for machine learning models," *arXiv preprint arXiv:2309.01440*, 2023.

[73] L. Risser, A. Picard, L. Hervier, and J.-M. Loubes, "A survey of identification and mitigation of machine learning algorithmic biases in image analysis," *arXiv preprint arXiv:2210.04491*, 2022.

[74] S. Georganos, T. Grippa, S. Vanhuysse, M. Lennert, M. Shimoni, S. Kalogirou, and E. Wolff, "Less is more: Optimizing classification performance through feature selection in a very-high-resolution remote sensing object-based urban application," *GIScience & remote sensing*, vol. 55, no. 2, pp. 221–242, 2018.

[75] Y. Chen, L. Ma, D. Yu, H. Zhang, K. Feng, X. Wang, and J. Song, "Comparison of feature selection methods for mapping soil organic matter in subtropical restored forests," *Ecological Indicators*, vol. 135, p. 108545, 2022.

[76] Y. Zhang, F. Shen, X. Sun, and K. Tan, "Marine big data-driven ensemble learning for estimating global phytoplankton group composition over two decades (1997–2020)," *Remote Sensing of Environment*, vol. 294, p. 113596, 2023.

[77] X.-Y. Tong, G.-S. Xia, Q. Lu, H. Shen, S. Li, S. You, and L. Zhang, "Land-cover classification with high-resolution remote sensing images

using transferable deep models," *Remote Sensing of Environment*, vol. 237, p. 111322, 2020.

[78] Z. Wang, L. Du, J. Mao, B. Liu, and D. Yang, "Sar target detection based on ssd with data augmentation and transfer learning," *IEEE Geoscience and Remote Sensing Letters*, vol. 16, no. 1, pp. 150–154, 2018.

[79] L. Zhang, M. Lan, J. Zhang, and D. Tao, "Stagewise unsupervised domain adaptation with adversarial self-training for road segmentation of remote-sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–13, 2021.

[80] A. Mumuni and F. Mumuni, "Data augmentation: A comprehensive survey of modern approaches," *Array*, p. 100258, 2022.

[81] H. Mansourifar, A. Moskowitz, B. Klingensmith, D. Mintas, and S. J. Simske, "Gan-based satellite imaging: A survey on techniques and applications," *IEEE Access*, 2022.

[82] P. Olofsson, G. M. Foody, M. Herold, S. V. Stehman, C. E. Woodcock, and M. A. Wulder, "Good practices for estimating area and assessing accuracy of land change," *Remote Sensing of Environment*, vol. 148, pp. 42–57, 2014.

[83] S. V. Stehman, "Sampling designs for accuracy assessment of land cover," *International Journal of Remote Sensing*, vol. 30, no. 20, pp. 5243–5272, 2009.

[84] C. Persello and L. Bruzzone, "Active and semisupervised learning for the classification of remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 52, no. 11, pp. 6937–6956, 2014.

[85] J. Li, D. Hong, L. Gao, J. Yao, K. Zheng, B. Zhang, and J. Chanussot, "Deep learning in multimodal remote sensing data fusion: A comprehensive review," *International Journal of Applied Earth Observation and Geoinformation*, vol. 112, p. 102926, 2022.

[86] C. Geiß, A. Rabuske, P. Aravena Pelizari, S. Bauer, and H. Taubenböck, "Selection of unlabeled source domains for domain adaptation in remote sensing," *Array*, vol. 15, p. 100233, 2022.

[87] J. Castillo-Navarro, B. Le Saux, A. Boulch, N. Audebert, and S. Lefèvre, "Semi-supervised semantic segmentation in earth observation: the minifrance suite, dataset analysis and multi-task network study," *Machine Learning*, vol. 111, pp. 3125–3160, 2022.

[88] C. M. Gevaert and M. Belgiu, "Assessing the generalization capability of deep learning networks for aerial image classification using landscape metrics," *International Journal of Applied Earth Observation and Geoinformation*, vol. 114, p. 103054, 2022.

[89] X. Zhu, W. Zhan, J. Zhou, X. Chen, Z. Liang, S. Xu, and J. Chen, "A novel framework to assess all-round performances of spatiotemporal fusion models," *Remote Sensing of Environment*, vol. 274, p. 113002, 2022.

[90] M. Rußwurm, L. H. Hughes, G. Pasquali, C. O. Dumitru, and D. Tuia, "Detection of settlements in tanzania and mozambique by many regional few-shot models," in *Proc. of the IEEE International Geoscience and Remote Sensing Symposium*, pp. 522–525, 2023.

[91] D. Tuia, F. Ratle, F. Pacifici, M. F. Kanevski, and W. J. Emery, "Active learning methods for remote sensing image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 47, no. 7, pp. 2218–2232, 2009.

[92] B. Demir, C. Persello, and L. Bruzzone, "Batch-mode active-learning methods for the interactive classification of remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 49, no. 3, pp. 1014–1031, 2010.

[93] M. Ferecatu and N. Boujemaa, "Interactive remote-sensing image retrieval using active relevance feedback," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 45, no. 4, pp. 818–826, 2007.

[94] Z. Fang, Y. Yang, Z. Li, W. Li, Y. Chen, L. Ma, and Q. Du, "Confident learning-based domain adaptation for hyperspectral image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–16, 2022.

[95] R. Gomes, P. Welinder, A. Krause, and P. Perona, "Crowdclustering," *Advances in neural information processing systems*, vol. 24, 2011.

[96] R. Roscher, W. Förstner, and B. Waske, "I2vm: Incremental import vector machines," *Image and Vision Computing*, vol. 30, no. 4-5, pp. 263–278, 2012.

[97] A. M. Wesley and T. C. Matisziw, "Methods for measuring geodiversity in large overhead imagery datasets," *IEEE Access*, vol. 9, pp. 100279–100293, 2021.

[98] E. Rolf, K. Klemmer, C. Robinson, and H. Kerner, "Mission critical–satellite data is a distinct modality in machine learning," *arXiv preprint arXiv:2402.01444*, 2024.

[99] J. Fowler, F. Waldner, and Z. Hochman, "All pixels are useful, but some are more useful: Efficient in situ data collection for crop-type mapping using sequential exploration methods," *International Journal of Applied Earth Observation and Geoinformation*, vol. 91, p. 102114, 2020.

[100] T. Burgert and B. Demir, "Estimating physical information consistency of channel data augmentation for remote sensing images," *arXiv preprint arXiv:2403.14547*, 2024.

[101] M. Willbo, A. Pirinen, J. Martinsson, E. L. Zec, O. Mogren, and M. Nilsson, "Impacts of color and texture distortions on earth observation data in deep learning," *arXiv preprint arXiv:2403.04385*, 2024.

[102] A. Kirsch, J. Van Amersfoort, and Y. Gal, "Batchbald: Efficient and diverse batch acquisition for deep bayesian active learning," *Advances in neural information processing systems*, vol. 32, 2019.

[103] D. Angluin and P. Laird, "Learning from noisy examples," *Machine Learning*, vol. 2, pp. 343–370, 1988.

[104] F. Bastani, P. Wolters, R. Gupta, J. Ferdinando, and A. Kembhavi, "Satlaspretrain: A large-scale dataset for remote sensing image understanding," in *Proc. of the IEEE/CVF International Conference on Computer Vision*, pp. 16772–16782, 2023.

[105] M. Rußwurm, S. J. Venkatesa, and D. Tuia, "Large-scale detection of marine debris in coastal areas with sentinel-2," *iScience*, 2023.

[106] D. Y. Zhang, Y. Huang, Y. Zhang, and D. Wang, "Crowd-assisted disaster scene assessment with human-ai interactive attention," in *Proc. of the AAAI Conference on Artificial Intelligence*, pp. 2717–2724, 2020.

[107] E. Saralioglu and O. Gungor, "Crowdsourcing in remote sensing: A review of applications and future directions," *IEEE Geoscience and Remote Sensing Magazine*, vol. 8, no. 4, pp. 89–110, 2020.

[108] H. Suresh and J. Guttag, "A framework for understanding sources of harm throughout the machine learning life cycle," in *Equity and access in algorithms, mechanisms, and optimization*, pp. 1–9, 2021.

[109] R. D. Cook, "Detection of influential observation in linear regression," *Technometrics*, vol. 19, no. 1, pp. 15–18, 1977.

[110] P. W. Koh and P. Liang, "Understanding black-box predictions via influence functions," in *Proc. of the International Conference on Machine Learning*, pp. 1885–1894, PMLR, 2017.

[111] Y. Chen, B. Li, H. Yu, P. Wu, and C. Miao, "Hydra: Hypergradient data relevance analysis for interpreting deep neural networks," in *Proc. of the AAAI Conference on Artificial Intelligence*, pp. 7081–7089, 2021.

[112] A. Ghorbani, M. Kim, and J. Zou, "A distributional framework for data valuation," in *Proc. of the International Conference on Machine Learning*, pp. 3535–3544, PMLR, 2020.

[113] T. Emmanuel, T. Maupong, D. Mpoeleng, T. Semong, B. Mphago, and O. Tabona, "A survey on missing data in machine learning," *Journal of Big Data*, vol. 8, no. 1, pp. 1–37, 2021.

[114] I. F. Ilyas and T. Rekatsinas, "Machine learning and data cleaning: Which serves the other?," *ACM Journal of Data and Information Quality (JDIQ)*, vol. 14, no. 3, pp. 1–11, 2022.

[115] F. Neutatz, B. Chen, Z. Abedjan, and E. Wu, "From cleaning before ml to cleaning for ml.," *IEEE Data Eng. Bull.*, vol. 44, no. 1, pp. 24–41, 2021.

[116] X. X. Zhu, J. Hu, C. Qiu, Y. Shi, J. Kang, L. Mou, H. Bagheri, M. Haberle, Y. Hua, R. Huang, *et al.*, "So2sat lcz42: A benchmark data

set for the classification of global local climate zones [software and data sets]," *IEEE Geoscience and Remote Sensing Magazine*, vol. 8, no. 3, pp. 76–89, 2020.

[117] J. M. Johnson and T. M. Khoshgoftaar, "Survey on deep learning with class imbalance," *Journal of Big Data*, vol. 6, no. 1, pp. 1–54, 2019.

[118] R. Roscher, B. Bohn, M. Duarte, and J. Garcke, "Explain it to me–facing remote sensing challenges in the bio-and geosciences with explainable machine learning," in *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. 3, pp. 817–824, Copernicus GmbH, 2020.

[119] X. Wang, Y. Chen, and W. Zhu, "A survey on curriculum learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 9, pp. 4555–4576, 2021.

[120] G. Varoquaux and O. Colliot, "Evaluating machine learning models and their diagnostic value," in *Machine Learning for Brain Disorders*, pp. 601–630, Springer, 2023.

[121] H. Meyer, C. Reudenbach, S. Wöllauer, and T. Nauss, "Importance of spatial predictor variable selection in machine learning applications–moving from data reproduction to spatial prediction," *Ecological Modelling*, vol. 411, p. 108815, 2019.

[122] J. Inglada, "Machine learning for land cover map production - follow-up on the tiselac challenge," 2018.

[123] A. M.-C. Wadoux, G. B. Heuvelink, S. De Bruin, and D. J. Brus, "Spatial cross-validation is not the right way to evaluate map accuracy," *Ecological Modelling*, vol. 457, p. 109692, 2021.

[124] G. M. Foody, "Sample size determination for image classification accuracy assessment and comparison," *International Journal of Remote Sensing*, vol. 30, no. 20, pp. 5273–5291, 2009.

[125] M. L. Blatchford, C. M. Mannaerts, and Y. Zeng, "Determining representative sample size for validation of continuous, large continental remote sensing data," *International Journal of Applied Earth Observation and Geoinformation*, vol. 94, p. 102235, 2021.

[126] R. G. Pontius Jr and M. Millones, "Death to kappa: birth of quantity disagreement and allocation disagreement for accuracy assessment," *International Journal of Remote Sensing*, vol. 32, no. 15, pp. 4407–4429, 2011.

[127] G. M. Foody, "Explaining the unsuitability of the kappa coefficient in the assessment and comparison of the accuracy of thematic maps obtained by image classification," *Remote Sensing of Environment*, vol. 239, p. 111630, 2020.

[128] E. Rolf, "Evaluation challenges for geospatial ml," in *Proc. of the ICLR Workshop on Machine Learning for Remote Sensing*, 2023.

[129] C. Robinson, K. Malkin, N. Jojic, H. Chen, R. Qin, C. Xiao, M. Schmitt, P. Ghamisi, R. Hänsch, and N. Yokoya, "Global land-cover mapping with weak supervision: Outcome of the 2020 ieee grss data fusion contest," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 14, pp. 3185–3199, 2021.

[130] M. Schmitt and Y.-L. Wu, "Remote sensing image classification with the sen12ms dataset," in *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. V-2-2021, pp. 101–106, 2021.

[131] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Transactions on knowledge and data engineering*, vol. 22, no. 10, pp. 1345–1359, 2010.

[132] M. Sugiyama, S. Nakajima, H. Kashima, P. Buenau, and M. Kawanabe, "Direct importance estimation with model selection and its application to covariate shift adaptation," *Advances in neural information processing systems*, vol. 20, 2007.

[133] J. Wen, R. Greiner, and D. Schuurmans, "Correcting covariate shift with the frank-wolfe algorithm," in *Proc. of the International Joint Conference on Artificial Intelligence*, 2015.

[134] A. de Mathelin, F. Deheeger, G. Richard, M. Mougeot, and N. Vayatis, "Adapt: Awesome domain adaptation python toolbox," *arXiv preprint arXiv:2107.03049*, 2021.

[135] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[136] W. R. Tobler, "A computer movie simulating urban growth in the detroit region," *Economic Geography*, vol. 46, pp. 234–240, 1970.

[137] S. Eyuboglu, M. Varma, K. K. Saab, J.-B. Delbrouck, C. Lee-Messer, J. Dunnmon, J. Zou, and C. Re, "Domino: Discovering systematic errors with cross-modal embeddings," in *Proc. of the International Conference on Learning Representations*, 2022.

[138] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "Simclr: A simple framework for contrastive learning of visual representations," in *Proc. of the International Conference on Learning Representations*, vol. 2, 2020.