Chapter 2
Analyzing and Improving the Quality and Fitness for Purpose of OpenStreetMap as Labels in Remote Sensing Applications

Moritz Schott, Adina Zell, Sven Lautenbach, Gencer Sumbul, Michael Schultz, Alexander Zipf, and Begüm Demir

Abstract OpenStreetMap (OSM) is a well-known example of volunteered geographic information. It has evolved to one of the most used geographic databases. As data quality of OSM is heterogeneous both in space and across different thematic domains, data quality assessment is of high importance for potential users of OSM data. As use cases differ with respect to their requirements, it is not data quality per se that is of interest for the user but fitness for purpose. We investigate the fitness for purpose of OSM to derive land-use and land-cover labels for remote sensing-based classification models. Therefore, we evaluated OSM land-use and land-cover information by two approaches: (1) assessment of OSM fitness for purpose for samples in relation to intrinsic data quality indicators at the scale of individual OSM objects and (2) assessment of OSM-derived multi-labels at the scale of remote sensing patches (1.22 × 1.22 km) in combination with deep learning approaches. The first approach was applied to 1000 randomly selected relevant OSM objects. The quality score for each OSM object in the samples was combined with a large set of intrinsic quality indicators (such as the experience of the mapper, the number of mappers in a region, and the number of edits made to the object) and auxiliary information about the location of the OSM object (such as the continent or the ecozone). Intrinsic indicators were derived by a newly developed tool

M. Schott (✉) · A. Zipf
Institute of Geography, GIScience, Heidelberg University, Heidelberg, Germany
e-mail: moritz.schott@uni-heidelberg.de; zipf@uni-heidelberg.de

A. Zell · G. Sumbul · B. Demir
Faculty of Electrical Engineering and Computer Science, Remote Sensing Image Analysis Group, Technische Universitát Berlin, Berlin, Germany
e-mail: adina.zell@campus.tu-berlin.de; gencer.suembuel@tu-berlin.de; demir@tu-berlin.de

S. Lautenbach
HeiGIT gGmbH at Heidelberg University, Heidelberg, Germany
e-mail: sven.lautenbach@heigit.org

M. Schultz
Institute of Geography, University of Tübingen, Tübingen, Germany
e-mail: michael.schultz@uni-tuebingen.de

D. Burghardt et al. (eds.), *Volunteered Geographic Information*,
https://doi.org/10.1007/978-3-031-35374-1_2

based on the OSHDB (OpenStreetMap History DataBase). Afterward, supervised and unsupervised shallow learning approaches were used to identify relationships between the indicators and the quality score. Overall, investigated OSM land-use objects were of high quality: both geometry and attribute information were mostly accurate. However, areas without any land-use information in OSM existed even in well-mapped areas such as Germany. The regression analysis at the level of the individual OSM objects revealed associations between intrinsic indicators, but also a strong variability. Even if more experienced mappers tend to produce higher quality and objects which underwent multiple edits tend to be of higher quality, an inexperienced mapper might map a perfect land-use polygon. This result indicates that it is hard to predict data quality of individual land-use objects purely on intrinsic data quality indicators. The second approach employed a label-noise robust deep learning method on remote sensing data with OSM labels. As the quality of the OSM labels was manually assessed beforehand, it was possible to control the amount of noise in the dataset during the experiment. The addition of artificial noise allowed for an even more fine-grained analysis on the effect of noise on prediction quality. The noise-tolerant deep learning method was capable to identify correct multi-labels even for situations with significant levels of noise added. The method was also used to identify areas where input labels were likely wrong. Thereby, it is possible to provide feedback to the OSM community as areas of concern can be flagged.

Keywords Volunteered geographic information · Data quality · Data analysis · Remote sensing · OpenStreetMap · Machine learning

2.1 Introduction

OpenStreetMap (OSM) has evolved to one of the most used geographic databases and is a prototype for volunteered geographic information (VGI). It is a major knowledge source for researchers, professionals, and the general public to answer geographically related questions. As a free and open community project, the OSM database can not only be edited but also used by any person with very limited restrictions such as internet access or usage citation. This open nature of the project enabled the establishment of a vibrant community that curates and maintains the projects' data and infrastructure, but also a growing ecosystem of tools that use or analyze the data (OpenStreetMap Contributors 2022a,b).

Recently, OSM has become a popular source of labeled data for the remote sensing community. However, spatial heterogeneous data quality provides challenges for the training of machine learning models. Frequently, OSM land-use and land-cover (LULC) data has thereby been taken at face value without critical reflection. And, while the quality and fitness for purpose of OSM data have been proven in many cases (e.g., Jokar Arsanjani et al. 2015; Fonte et al. 2015), these analyses have also unveiled quality variations, e.g., between rural and urban regions. The quality of OSM can thus be assumed to be generally high, but remains unknown for a specific use case. It is therefore of importance to develop both tools that are capable of quantifying data quality of LULC information in OSM and approaches that are capable of dealing with the noise potentially present in OSM.

The IDEAL-VGI project investigated the fitness for purpose of OSM to derive LULC labels for remote sensing-based classification models by two approaches: (1) assessment of OSM fitness for purpose for samples in relation to intrinsic data quality indicators at the scale of individual OSM objects and (2) assessment of OSM-derived multi-labels at the scale of remote sensing patches (1.22×1.22 km) in combination with deep learning methods.

2.2 Intrinsic Data Quality Analysis for OSM LULC Objects

One of the most prominent analysis topics in OSM-related research is data quality that has been covered in theory (see, e.g., Barron et al. 2014; Senaratne et al. 2017) as well as in many practical studies (e.g., Jokar Arsanjani et al. 2015; Brückner et al. 2021). The topic of data quality is of concern for many studies working with volunteered geographic information—Chap. 1, for example, deals with data quality in OSM and Wikidata. Senaratne et al. (2017) characterize analyses into extrinsic metrics, where OSM is compared to another dataset, and intrinsic indicators, where metrics are calculated from the data itself. Semi-intrinsic (or semi-extrinsic) metrics use auxiliary information to assess the quality of OSM—population density can, for example, be used to assess the completeness of buildings in OSM, as population density and number of buildings are related. The quality gold standard has frequently been defined for extrinsic metrics through an external dataset of higher or known quality and standards. However, external datasets of high quality—including high up-to-dateness—are frequently not available. Therefore, intrinsic data quality indicators have frequently been used (Barron et al. 2014). These try to capture data quality aspects based on the history of OSM data itself, such as the number of edits to an object. Although OSM objects can be viewed individually, they are always embedded in a larger context of surrounding OSM objects, communities of contributors, and other classification systems, such as biomes or socioeconomic factors. Comparing contributions and communities for selected cities, (Neis et al. 2013), e.g., found a positive correlation between contributor density and gross national product per capita and showed that community sizes vary between Europe and other regions. In 2021, Schott et al. (2021) described "digital" and "physical locations" in which an OSM object is located. These "locations" consist of, intrinsic, OSM-specific measures such as density and diversity of elements, but also include—semi-intrinsic—aspects of economic status, culture, and population density to describe the surrounding of an object. Such information provides potentially relevant information to help characterize and predict data quality of OSM objects.

LULC information in OSM is a challenging topic. On the one hand, this information provides the background for all other data rendered on the central map.

It can highly benefit from local input, survey mapping, and (live) updates. On the other hand, this information has a difficult position within the OSM ecosystem. While the routing and the building of communities are prominent, LULC is not so frequently mentioned in the ecosystems' communication platforms. LULC information can also be quite cumbersome or even difficult to map, e.g., due to natural ambiguity. The growing tagging scheme provides a collection of sometimes ambiguous or overlapping tag definitions that are not fully compatible with any official LULC legend definition (Fonte et al. 2016). Furthermore, the data is highly shaped by national preferences and imports.

2.2.1 OSM Element Vectorization: Intrinsic and Semi-intrinsic Data Quality Indicators

The OSM element vectorization tool (OEV, Schott et al. 2022) has been developed to ease access to intrinsic and semi-intrinsic indicators, with a specific focus on LULC feature classes. The tool[1] provides access to currently 31 indicators at the level of single OSM objects (c.f. Table 2.1), which cover aspects concerning the element itself, surrounding objects, and the editors of the object.

The usability of the tool was proven on the use case of LULC polygons. One thousand out of the globally existing 62.9 million LULC elements were randomly sampled on 2022-01-01. Only polygonal objects with at least one of the LULC defining tags were considered. These elements' IDs were then fed to the tool to extract the data and calculate the described metrics from Table 2.1. These metrics were used in a cluster analysis to identify structures in the OSM LULC objects. Furthermore, we tested three hypotheses on the triangular relation between the size of OSM objects, their age, and their location in terms of population density. We hypothesized that a general mapping order exists where the OSM community first concentrates on or arises from urban areas before moving to rural areas. This was tested by the hypotheses 1 (H_1): *There is a positive correlation between the object age and the population density*. Second, we tested the hypotheses that areas with higher population density are more fragmented and therefore exhibit smaller elements, while areas with low population density, such as forest, are often larger objects: (H_2) *there is a negative correlation between the object size and the population density*. Third, we tested the effect between the OSM LULC objects' age and population density, assuming a non-significant correlation. This was based on two opposing assumptions: Large geographical entities may be mapped first, and regions may be first coarsely drafted before adding details. This would lead to old objects being of larger size. Yet, hypotheses 1 and 2 contradict this tendency: according to H_1 and H_2, younger objects would be in areas with less population density and therefore tend to be larger. All three hypotheses were tested separately

[1] https://oev.geog.uni-heidelberg.de/.

Table 2.1 Intrinsic and semi-intrinsic indicators calculated by the OEV tool. The indicators are grouped in the categories of semantic, geometric, surrounding, OSM surrounding, temporal, mapper, mapping process, and linting tools

Indicator	Description	Logical link to LULC quality
osm_type	Type of the OSM object (node, way, relation)	Relations in OSM are more complex and need higher skill to be mapped but also are more error-prone
primtag	Primary tag of the object	Some tags may be more difficult to map, ambiguous, less established, or less documented than others
corine_class	Aggregation of primary tags at CORINE level 2	
invalidity	Gravity of geometric invalidity	Invalid geometries describe bad data
Detail	How detailed is the object drawn?	
Complexity	How complex is the geometry?	The more detailed/less course an object is drawn, the higher was the effort by the mapper and a higher quality can be assumed
obj_size	Size of the object	The smaller the element, the better (up to a certain point)
continent	The country or continent an element is located on	Regions have different communities, physical geographic contexts, etc. that influence quality
pop5k	Population density in the objects' surrounding	Findings are that rural areas often have lower quality; how many potential local mappers are available and what features are plausible given the population density?
hdi	Economic situation of a region	The economic status influences mapping (leisure time, equipment, etc.)
biome	The "physical appearance" of the objects' surrounding	The biome influences the probability/validity of certain objects
rs_variance	How homogeneous is the object in remote sensing imagery?	Homogeneous reflectance -> homogeneous area -> object well drawn (depending on object type)

(continued)

Table 2.1 (continued)

Indicator	Description	Logical link to LULC quality
overlap	How much does the element overlap with elements in the surrounding?	Less overlap -> better geometry
shared_ border	What part of the object's border is shared with surrounding objects?	If the object is well placed, other objects might be closely attached to it. As LULC is optimally space filling, a well-mapped area would be represented by connected LULC objects
surrounding_ mapped	How well is the area mapped in general?	Well-mapped regions might indicate good quality
surrounding_ diversity	How many different primary tags relevant for LULC are present in the surrounding?	Diversity of mapped LULC objects might indicate high quality—if real-world LULC is characterized by some diversity. Often some LULC classes are mapped first—if surrounding diversity is high, this could indicate that the mapping has progressed beyond this point
surrounding_ obj_dens	How many objects are there in general in the area?	Well-mapped areas have many objects (in relation to population density, etc.)
how_many_ eyes	How many contributors were involved in this area?	Does this area even deserve the "crowdsourced" tag?
oqt_mapping_ saturation	Estimated area mapping completeness?	Saturated areas may have better quality as more effort by mappers can be put into improvement instead of the creation of missing objects

outofdateness	How outdated is the object?	Recently changed objects are potentially more up to date
object_age	How old is the object?	Old objects are either outdated and were mapped with superseded mapping practices or had much time to mature with many small improvements
changes_ per_year	How frequently has the object been changed/updated?	Objects that are regularly changed are well maintained
user_mean_exp	How experienced were the users that changed this object?	More experienced users may have a higher probability to make changes that improve quality
user_remoteness	How far away were other edits done by the mapper(s) of the object?	Local users are more familiar with the region and should create better data
user_diversity	How specialized is the user?	Diverse users may be general experts, but specialized users may be experts on this topic
data_source	What sources were given for the changes?	Certain sources may be more reliable
cs_editor	Editor/editing software used	Some editors might have bugs or are not suited for certain editing
cs_median_area	Was the object edited by local edits or large-scale edits?	Changesets that span large areas hint toward bad quality, imports, or unspecific changes
validator_issued_ density	Density of issues reported by linting tools	Less errors indicate a well-maintained region and good quality
n_osm_notes	How many OSM notes are there?	Many notes either show bad data or an active community
osm_cha_flags	What OSMCha flags were added to the changesets?	Many flags are related to suspicious changesets or edits

using Kendall's τ (Hollander and Wolfe 1973), with the p-values adjusted for multiple testing (Benjamini and Hochberg 1995).

Results of this first exploratory analysis provided interesting insights into the complex and multifaceted structure of OSM land-use objects and its relation to OSM mapping communities. The main hypotheses regarding the mapping order (H_1) could not be confirmed. In fact, the estimated correlation was slightly negative, meaning that for the used sample, objects in urban areas were younger than in rural areas. Yet, this does not imply that this mapping order does not exist in certain subregions. In addition, the age of an object is a fragile metric that highly depends on the mapping style of local mappers. Mappers frequently decide to delete and redraw elements instead of changing the original object, especially if the object was only a coarse approximation. This "resets" the object age, meaning that urban areas may have a high share of young objects because they are still actively mapped and maintained, even though they started their map appearance relatively early. H_3 was equally confirmed, but only after p-value correction (p-value $= 0.14$). Regional specialities may exist in this aspect and need further investigation.

The negative correlation between the object size and the population density (H_2) was confirmed with a p-value<0.01 though the τ was only -0.096 implying a small effect. At the global scale, many influencing factors may overlap or intervene with each other, hindering the extraction of single detailed effects. For the example at hand, we can assume that there are multiple regional communities or active mappers with individual mapping styles. The mapping detail in urban or rural regions will therefore be linked to these and other factors as well, not only the population density. Population density itself may not be generalizable on a global scale. The same level of fragmentation, meaning object size distribution, may be reached at different population density values, depending, e.g., on the continent.

The cluster analysis revealed interesting aspects, as some clusters could be associated with imports. Especially, a large import of North American lakes could be separated. This element group made up a considerable share of the global data and must therefore be taken into account when analyzing or describing the global dataset.

One thousand LULC objects were manually checked against high-resolution imagery. A combined quality score was assigned based on the thematic and the geometric correctness of the object. A quantile random forest was used to identify relationships between the data quality score and the 31 indicators calculated by the OEV tool.

While the overall quality of the model was intermediate, we were able to identify a series of interesting relationships between the indicators and the quality of the land-use objects based on the visual inspection of partial dependency plots. The most important features in the model (c.f. Fig. 2.1) were the size of the OSM object, contributor characteristics (such as experience and remoteness), rare OSM tags, and regional OSM mapping aspects (e.g., number of OSM objects in the surrounding of the object).

Element size had by far the highest feature importance. However, the effect on data quality of the OSM object had no clear direction. The indicator had to

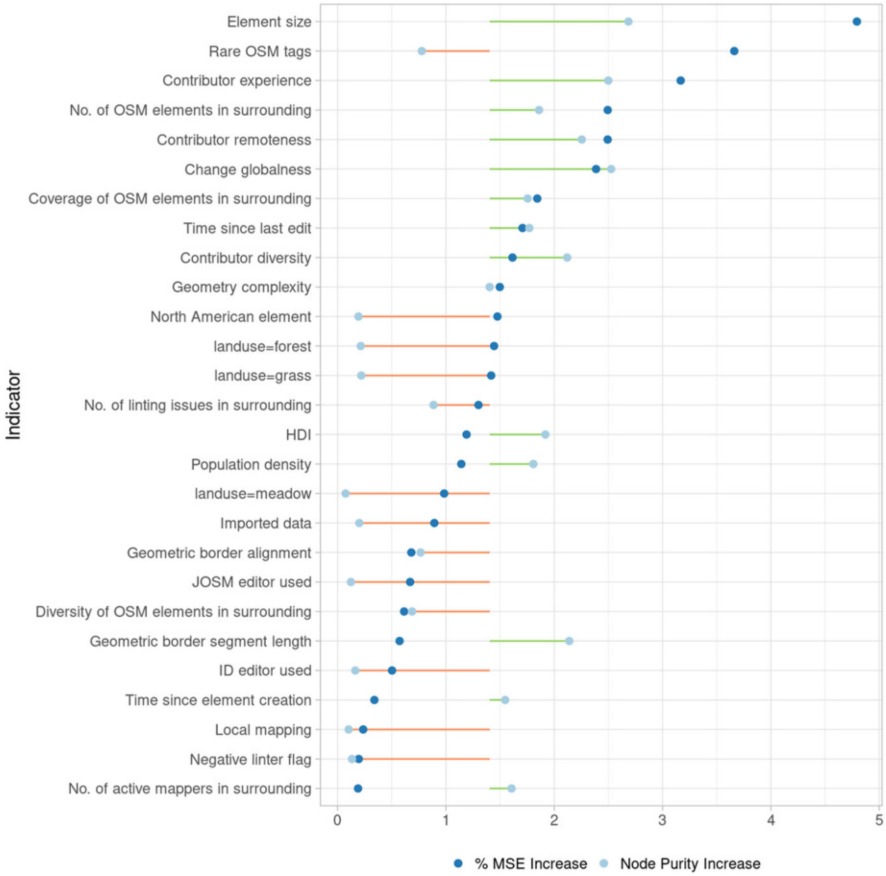

Fig. 2.1 Feature importance in relation to data quality. The importance was derived based on a quantile random forest for 1000 randomly selected OSM objects. The features are sorted by the percentage increase in squared mean error if the feature would be dropped. In addition, node purity is provided as a second feature importance indicator. To ease interpretation, the second indicator is displayed together with its position relative to the median node purity value across all selected features

be interpreted in combination with other indicators such as the primary tag (e.g., *landuse* or *natural*) as some land-use tags were characterized by big objects (e.g., forest) and others by small objects (e.g., urban grass). Objects with rare primary tags indicated, in general, a lower object quality—presumably as these tags are less well established and possibly poorly defined and thereby harder to map consistently. The effect of contributor experience showed a multimodal distribution: contributors with very little experience (newcomers) were associated with objects of medium quality and contributors with medium experience (stable mappers) with high quality of the land-use objects. Interestingly, highly experienced mappers were associated with poor quality of the land-use object. One possible explanation is that these extraordinary active users might represent bots or importers. One has, however, to keep in mind that only a few contributors fell into this category, which led to a low statistical power.

With respect to the number of elements surrounding an OSM LULC object, areas with more elements were—expectedly—better mapped. The larger the share of the surrounding of an OSM object mapped by LULC, the higher the quality of the OSM object. However, shares above 100% expectedly were associated with a lower quality of the OSM object. These regions are characterized by areas mapped using highly overlapping polygons, which typically indicate mapping errors. OSM edits are contributed as changesets. Larger changesets (more elements, often from different regions) were associated with lower quality of the OSM land-use objects. This is in line with the expectation that local, concise, and coherent edits are better. With respect to the primary tag (the LULC class), the model indicated differences in the quality of some classes: while forest objects were of higher quality, grass-dominated LULC classes were of lower quality. This could be explained by the clear distinction of forest from other LULC classes and the diversity of tags used to characterize different grass-dominated LULC classes. LULC objects in North America had a tendency for lower quality, presumably due to the large imports in this region. Besides that, LULC objects mapped in regions with higher Human Development Index or higher population density were associated with better data quality. This presumably reflects the larger OSM community in the Global North, especially in countries such as Austria, Germany, or France, as well as the higher number of potential mappers available in urban areas compared to the countryside. With respect to out-of-dateness, complex interactions were identified; however, generally recently changed objects were associated with higher quality. Except for newcomers, lower user diversity—i.e., users focusing on one aspect of OSM—was associated with higher data quality.

2.3 Label Noise Robust Deep Learning for Remote Sensing Data with OSM Tags

2.3.1 OSM as the Source of Training RS Image Labels in ML

Supervised ML methods have attracted great attention for Earth observation applications on ever-growing RS image archives. Due to their capability to automatically model higher-level RS image semantics in large scale, they are applied to many problems in RS such as multi-label image classification and land-cover map generation. These methods generally require the availability of a high quantity of annotated training RS images. However, the manual collection of RS image annotations by domain experts for a large amount of data can be time-consuming, complex, and costly. Accordingly, the use of volunteered geographic information as crowdsourced data such as OSM to automatically derive annotated training data has been drawing significant attention in RS. As an example, in (Kaiser et al. 2017; Wan et al. 2017; Comandur and Kak 2021), it is shown that the direct use of OSM tags as pixel-level land-use class labels is useful for the automatic map generation of RS images through support vector machines and convolutional neural networks (CNNs). In (Li et al. 2020), OSM tags are utilized as scene-level class labels of training RS images to automatically predict land-use/land-cover classes of RS image scenes through CNNs. In (Audebert et al. 2017), OSM data is also utilized as an auxiliary information source by fusing it with optical data from very-high-resolution satellite imagery through dual-stream CNNs. In (Lin et al. 2022), an active learning strategy is introduced to partially annotate RS images with salient multi-labels based on OSM tags. In this study, an adaptive temperature-associated model is also proposed to apply multi-label RS image classification by utilizing partially annotated training data and automatically assigning missing labels to training images during training.

Thanks to the publicly available OSM database, collection of RS image annotations for a high quantity of training data to be utilized for ML methods can be achieved at lower costs. However, OSM tags can be outdated regarding RS images due to possible changes on the ground; or there can be annotation errors. Accordingly, using OSM tags as the source of training image annotations may increase the chance of including noisy labels in training data of ML methods. As an example, for multi-label image annotations of RS images, two types of noise can exist. Noise can be associated with missing labels or wrong labels. A missing label means that although a land-use/land-cover class exists in an RS image, the corresponding class label is not assigned. A wrong label means that although a class label is assigned to RS image, the corresponding class is not present in the image.

2.3.2 Label Noise Robust ML Methods

When a ML model is trained on noisy training data, there is a risk of overfitting of the model parameters to noisy labels and thus suboptimal inference performance. To this end, a few methods are presented in RS to improve the robustness of ML models toward noisy labels in training data. As an example, in (Zhang et al. 2020a), a noisy label knowledge distillation method is introduced for single-label RS image classification problems to leverage the knowledge learned through a teacher model on images with noisy labels for a student model. In this method, two CNNs are employed as a teacher-student framework, while a clean and trustworthy subset of a training set is assumed to be available for the student CNN. In (Aksoy et al. 2022), a collaborative learning framework is proposed to identify and exclude images with noisy multi-labels during training. To this end, it employs two CNNs operating collaboratively, while they are forced to characterize distinct image representations and to produce similar predictions. In (Burgert et al. 2022), the effects of the abovementioned label noise types in multi-label RS image classification problems are investigated, while different single-label noise robust methods are integrated to multi-label classification problems in RS. In (Dong et al. 2022), for land-cover map generation through semantic segmentation, an online noise correction approach is introduced to detect and correct pixel-level noisy labels via information entropy at the early stage of model training and thus to continue training with corrected labels. Although all these methods are potentially effective, the development of label noise robust ML methods when the OSM tags are utilized as the source of image annotations has not yet been investigated in RS literature.

It is worth mentioning that learning the parameters of ML models under noisy labels in training data has been studied more extensively in computer vision (CV) literature than RS. Recent research directions in CV can be grouped into the development of (1) deep neural network (DNN) architectures, (2) ML loss functions, (3) regularization strategies for training ML models, and (4) training sample selection and label adjustment techniques for single-label image classification problems. The first set of methods are concentrated on the development of DNN architectures designed for training data with noisy labels. For example, a contrastive-additive noise network is introduced in (Yao et al. 2019) to model trustworthiness of noisy training labels. This network consists of a probabilistic latent variable model as a contrastive layer in order to measure the quality of annotations and an additive layer to aggregate the class predictions and noisy labels. The second set of methods is mostly focused on the development of ML loss functions, which embody robust characteristics toward noisy labels. As an example, in (Ridnik et al. 2021), an asymmetric loss function is proposed to dynamically decrease the weights of negative classes in multi-labels. This allows to decrease the effect of images with missing labels on ML model parameter updates during training. The third set of methods are concentrated on regularizing the whole ML model training to prevent overfitting of model parameters to noisy labels. For instance, a regularization term is integrated into the cross-entropy loss function in (Liu et al. 2020) to utilize the class predictions from an early stage of ML model training to prevent the memorization of noisy labels. The fourth set of methods aim to first select images with correct labels or adjust noisy labels and then to learn through samples with correct labels. As an example, a joint training with co-regularization approach is introduced in (Wei et al. 2020) to employ collaborative learning of two CNNs for the selection of correct labels by an agreement strategy.

2.3.3 Proposed Methods

Due to the public availability of OSM, RS images can be automatically associated with multiple land-use/land-cover classes (i.e., multi-labels) by using OSM tags. This allows to create large training sets for deep learning (DL)-based multi-label RS image classification methods at lower costs. Let $\mathcal{X}=\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_M\}$ be an RS image archive that includes M images, where $\boldsymbol{x}_k$ is the kth image in the archive. We assume that a training set $\mathcal{T} = \{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i=1}^{D}$ is available. Each training image

$\boldsymbol{x}_i$ is associated with a set of class labels $\boldsymbol{y}_i \in \{0, 1\}^S$ based on the corresponding OSM tags, where S is the total number of classes. Let $\phi : \theta, \mathcal{X} \mapsto \hat{\mathcal{Y}}$ be any type of convolutional neural network (CNN) that generates the multi-label $\hat{\boldsymbol{y}}_k$ of an image $\boldsymbol{x}_k \in \mathcal{X}$. Training ϕ on $\mathcal{T}$, which may include noisy labels due to noisy OSM tags, can lead to learning suboptimal model parameters θ and inaccurate inference performance, as discussed in the previous sections.

To address this issue, we aim to first automatically detect noisy OSM tags based on the CNN ϕ trained on $\mathcal{T}$ and then adjust training labels associated with noisy OSM tags for label noise robust learning of the CNN model parameters θ.

2.3.3.1 Noisy OSM Tag Detection

Region-based RS image representations combining both local information and the related spatial organization of land-use/land-cover classes are important for the accurate detection of noisy OSM tags. However, multi-label RS image predictions $\hat{\mathcal{Y}}$ of the considered CNN ϕ do not provide spatial information regarding the class location.

Accordingly, we employed class activation maps (CAM) introduced in (Zhou et al. 2016) since they are capable of deriving the regions most relevant for a given class with respect to the DL model trained for image classification. Let $\mathcal{F}_i \in \mathbb{R}^{CxWxH}$ be a set of feature maps for an image $\boldsymbol{x}_i$ obtained from the last convolutional layer of the CNN backbone where C, H, and W represent the number of channels, height, and width of the feature maps, respectively. CAMs associated with $\boldsymbol{x}_i$ can be obtained by applying a 1x1 convolutional layer, which takes the feature maps $\mathcal{F}_i$ of $\boldsymbol{x}_i$ as input and produces a set of feature maps $\mathcal{A}_i \in \mathbb{R}^{SxWxH}$. The sth feature map $\mathcal{A}_i^s \in \mathbb{R}^{WxH}$ is the localization map associated with a class s, which can be obtained as follows:

$$\mathcal{A}_i^s = \sum_{c=1}^{C} w_s^c \mathcal{F}_i^c \tag{2.1}$$

where w_s^c is the weight of importance for the cth feature map $\mathcal{F}_i^c$ regarding the sth class. The obtained CAMs are forwarded through a global average pooling (GAP) layer to obtain multi-label class predictions. However, multi-label classification models from which CAMs can be derived are trained only to identify the presence of a given class within the image. Thus, CAMs tend to focus only on the most discriminative features within the image, leading to the incomplete coverage of the target class within the image (Zhang et al. 2020b).

Self-enhancement maps (SEMs) introduced in (Zhang et al. 2020b) address this issue and improve the localization maps derived from CAMs by including the similarity of feature maps in the localization map calculation. This is achieved by first defining seed coordinates, which are the image regions with the largest activation values on CAMs for a given class. Then, a similarity map is created for each seed point based on the cosine similarity between the seed point feature vector and all other feature vectors. For a given class s of an image $\boldsymbol{x}_i$, the final class activation map $\mathcal{E}_i^s$ is obtained by taking the maximum value at each pixel across the similarity maps. After obtaining all the class activation maps, we generate a class prototype P^s for the class s by following (Lee et al. 2018). This is achieved by averaging all the feature maps, which are extracted from $\mathcal{T}$ and associated with the class s. Class prototypes allow us to obtain more accurate class predictions based on spatial information regarding the location of image classes and the corresponding region-based image representations. Accordingly, to define whether the class s is present in the image $\boldsymbol{x}_i$, the extracted features of the image for the class $\mathcal{F}_i^s$ are compared with the corresponding class prototype based on their cosine similarity as follows:

$$\hat{\boldsymbol{y}}_i^s = \begin{cases} 1, & \cos(P^s, \mathcal{F}_i^s) > 0.5 \\ 0, & \text{otherwise} \end{cases} \tag{2.2}$$

To detect if x_i is associated with noisy labels, we compare the class predictions $\hat{\boldsymbol{y}}_i$ with the associated OSM tags. If the CNN model predicts a class which is not in the list of class labels derived from OSM tags, it is assumed to be a missing class. This missing class can be localized through SEMs. If the CNN model does not predict a class, but it is in the list of class labels derived from OSM tags, it is assumed to be a wrong class label. It is noted that automatically defining noisy OSM tags and the localization of missing classes via SEMs allow providing feedback to the OSM community. Such feedback together with further investigations in the OSM community can lead to correcting noisy OSM tags by human mappers.

2.3.3.2 Label Noise Robust Multi-label RS Image Classification

It is worth noting that the abovementioned method for the detection of noisy OSM tags relies on the model parameters θ of the considered CNN, which is obtained through training on $\mathcal{T}$. If a small trustworthy subset $C \in \mathcal{T}$ of the training set is available, this method can also be used to automatically find training images associated with noisy labels.

To this end, we divide the whole learning procedure into two stages. In the first stage, θ is learned by training ϕ only on C. After this stage is finalized, we first automatically divide the rest of the training set $\mathcal{T} \setminus C$ into training images with noisy labels $\mathcal{N}$ and training images with correct labels $\mathcal{L}$ (i.e., $\mathcal{N} \cup \mathcal{L} = \mathcal{T} \setminus C$). Then, class labels associated with each image in $\mathcal{N}$ are automatically corrected based on (2.2) leading to training images with corrected labels $\mathcal{N}^*$. This leads to automatically correcting noisy labels in $\mathcal{T} \setminus C$ derived from OSM tags. Then, the training set of the second stage $\mathcal{T}^*$ is formed by combining $\mathcal{L}$, C, and $\mathcal{N}^*$.

In the second stage, all the model parameters of ϕ are fine-tuned on $\mathcal{T}^*$. Thanks to the first stage, noisy labels included in the training set of this stage are significantly reduced. This allows to overcome overfitting on noisy labels of the whole training set in the second stage. Due to this two-stage learning of the model parameters, abundant training RS images annotated with OSM tags can be facilitated for label noise robust learning of multi-label RS image classification through CNNs.

2.3.4 Results and Discussion

In this subsection, we first describe the considered dataset and the experimental setup and then provide our analysis of the experimental results.

2.3.4.1 Dataset Description and Experimental Setup

To conduct experiments, we selected a Sentinel-2 tile acquired over South-West Germany including parts of France on 2021-06-13. This region spans from the Palatinate Forest in the west to the Odenwald in the east and includes large forested areas as well as areas dominated by agriculture or by built-up areas. This tile was divided into 81001.22 × 1.22-km-sized image patches. Each image patch is annotated with multi-labels based on the presence or absence of four major land-use classes that are defined with OSM tags (c.f. Table 2.2). While assigning the labels, small OSM objects were filtered (see Table 2.2 for thresholds used). The resulting labels of 910 image patches were manually validated against Sentinel-2 imagery.

Table 2.2 OSM land-use classes used for the multi-label image classification

Class	Description and filter	OSM tags
Water bodies	Continuous 0.2 ha of non-intermittent surface water; smaller ponds and all pools were not considered	landuse=reservoir, natural=water, waterway=dock, waterway=riverbank
Forests	Continuous 0.5 ha closed tree cover; smaller tree groups were not considered	landuse=forest, natural=wood
Agricultural areas	Continuous 0.5 ha meadow; arable land or vineyards, non-agricultural areas (parks, etc.), and smaller isolated elements were not considered assuming they are non-agricultural gardens or similar	landuse=farmland, landuse=meadow, landuse=vineyard
Built-up areas	Continuous 0.5 ha containing mostly impermeable features (buildings, roads, etc.); single isolated buildings are not considered, and large permeable objects like parks or sports grounds are not part of the built-up area	landuse=civic_admin, landuse=commercial, landuse=depot, landuse=education, landuse=farmyard, landuse=garages, landuse=industrial, landuse=residential, landuse=retail

Table 2.3 Multi-label image classification results in terms of mean average precision (mAP) obtained by the direct use of OSM tags (OSM) with different values of synthetic label noise rate (SLNR) of test set and DeepLabV3+ with different values of SLNR of training set

Method	SLNR	mAP (%)
OSM	0%	94.2
	10%	89.4
	20%	83.1
	30%	77.3
	40%	72.9
DeepLabV3+	0%	99.2
	20%	96.8
	40%	70.9
	60%	66.6
	80%	60.4

The OSM quality and completeness in the region were high (c.f. Table 2.3). OSM-based multi-label assignments had a mean average precision of 94.2%. The patches were clustered with respect to the correct assignment of the multi-labels: Clusters of correct OSM data were often due to monotonous landscapes, e.g., the Palatinate Forest. Clusters of flawed data were often due to missing data, e.g., in the region around Kaiserslautern. To perform experiments, 200 manually labeled patches were used as the test set, while the rest of the image patches were utilized as the training set.

In the experiments, we utilized the DeepLabv3+ (Chen et al. 2018) CNN architecture as the DL model. It is worth noting that DeepLabv3+ is originally designed for semantic segmentation problems. We replaced its semantic segmentation head with a fully connected layer followed by a GAP layer that forms the multi-label classification head with four output classes. We trained DeepLabv3+ for 20 epochs with Adam optimizer and the initial learning rate of 0.001. For the proposed label noise robust learning method, the same number of training epochs is used for each of the first and second stages. Experimental results are provided in terms of micro mean average precision (mAP) scores and noise detection accuracies.

We conducted experiments to (i) compare the considered DL model with the direct use of OSM tags for multi-label RS image classification, (ii) analyze the effectiveness of the proposed label noise detection method, and (iii) assess the effectiveness of the proposed label noise robust learning method. For the proposed label noise robust learning method, which requires the availability of a small trustworthy subset of the training set, we included the manually labeled image patches to the training set. However, for the comparison between the DL model and the OSM tags, only non-verified training data was utilized. To assess the robustness of the CNN model toward label noise and to detect noisy samples, we injected synthetic label noise to the training and test sets at different percentages (which were 20%, 40%, 60%, and 80% for the training set and 10%, 20%, 30%, and 40% for the test set) by following (Burgert et al. 2022).

2.3.4.2 Comparison Between Direct Use of OSM Tags and DL-Based Multi-label Image Classification

In this subsection, we assess the effectiveness of the considered DL model (DeepLabV3+) compared to the direct use of OSM tags for multi-label RS image classification. To this end, the model parameters of the DL model were learned on abundant non-verified training data without considering label noise robust learning. Table 2.3 shows the corresponding results in terms of mAP values, when the different values of synthetic label noise rate (SLNR) were applied to the training set of DeepLabV3+ and the OSM tags. One can observe from the table that when SLNR equals to 0% for training and test sets, DeepLabV3+ achieves 5% higher mAP values compared to directly using OSM tags. Even when 20% label noise is synthetically added to the training set of the CNN model, it is still capable of achieving higher results compared to OSM when SLNR value equals to 0%. It is worth mentioning that directly using OSM of such low quality leads to missing or wrong classes. It can be seen from the table that when synthetic noise is added to OSM tags, its multi-label image classification performance is significantly reduced. As an example, when SLNR value is increased to 20% from 0%, multi-label image classification performance of OSM is reduced by more than 10%. These results show the effectiveness of using OSM as a training source of CNN models compared to directly using OSM tags for multi-label image classification. This is relevant because preliminary OSM data analyses may not be able to confidently identify such malicious areas of bad quality.

It is worth noting that further increasing the SLNR value of the training set of DeepLabV3+ significantly reduces multi-label image classification performance. This is due to the fact that when a training set of a DL model includes a higher rate of noisy labels, the model parameters are overfitted on noisy labels that lead to suboptimal learning of multi-label image classification. Figure 2.2 shows the self-enhancement maps (SEMs) of an RS image obtained on DeepLabV3+ trained under different values of SLNR. One can see from the figure that as the SLNR value of the training set increases, the capability of CNN model to characterize the semantic content of the image reduces due to noisy labels.

2.3.4.3 Label Noise Detection

In this subsection, we assess the effectiveness of the proposed label noise detection method when different rates of synthetic label noise are applied to the test set. We also analyze the effect of the level of label noise (which is present in the training data) on our method. Table 2.4 shows the corresponding label noise detection accuracies obtained on DeepLabV3+ trained with abundant non-verified training data at different SLNR values and a small data, which is verified in terms of label noise. One can observe from the table that when SLNR equals to 0%, our label noise detection method, which is applied to DeepLabV3+ and trained on abundant data, achieves the highest label noise detection accuracies. For example, when synthetic

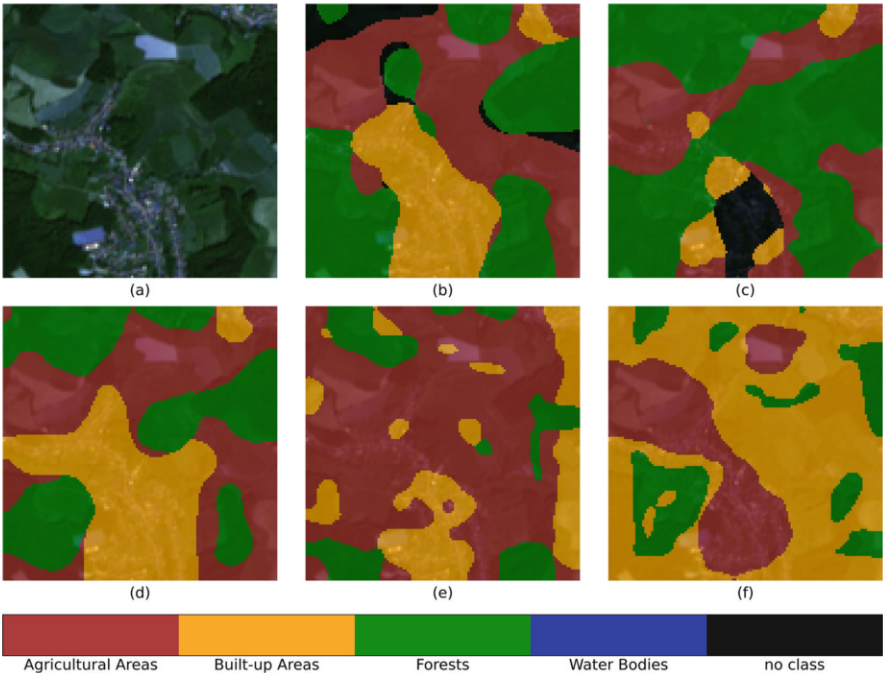

Fig. 2.2 (**a**) An example of RS image and its self-enhancement maps obtained on DeepLabV3+ trained under synthetic label noise rates (**b**) 0%, (**c**) 10%, (**d**) 20%, (**e**) 30%, and (**f**) 40%

Table 2.4 Label noise detection results in terms of accuracy (%) obtained by the proposed label noise detection method applied to the DeepLabV3+ trained with abundant non-verified data at different values of synthetic label noise rate (SLNR) and small verified data

Training Set	**SLNR (Test Set)**				
	0%	10%	20%	30%	40%
Abundant non-verified data (SLNR = 0%)	80.0	88.5	87.0	89.0	94.0
Abundant non-verified data (SLNR = 20%)	63.0	72.5	79.0	82.5	86.5
Abundant non-verified data (SLNR = 40%)	0.5	30.0	50.5	64.5	71.5
Abundant non-verified data (SLNR = 60%)	0.0	31.0	55.5	65.5	74.0
Abundant non-verified data (SLNR = 80%)	0.0	31.5	55.5	65.0	74.0
Small verified data	45.0	58.5	67.0	78.5	84.0

label noise is applied to the test set with 40%, our label noise detection method is capable of achieving 94% accuracy. As the SLNR value increases on abundant training data, label noise detection accuracy of our method decreases. This is in line with our conclusion from the previous subsection about the effect of label noise level in training data. In greater details, when SLNR value reaches to 40% for abundant training data, noise detection accuracy of our method decreases by more than 50% compared to SLNR = 0%. When a small verified training data with the size of 10% of the whole training set is used for DeepLabV3+, our noise detection method achieves higher accuracies compared to abundant training data with SLNR $\geq$ 40%. These results show that our label noise detection method is capable of effectively detecting noisy labels without requiring a small verified data when the amount of label noise in the training data is small. However, if the level of label noise in training data is greater than a certain extent, our method requires the availability of a small trustworthy subset of the training set for accurate label noise detection.

2.3.4.4 Label Noise Robust Multi-label Image Classification

In this subsection, we compare the proposed label noise robust learning method with the standard learning procedure, in which label noise of a training set is not considered during training. Table 2.5 shows the corresponding multi-label RS image classification scores when synthetic label noise is injected to the training set at different values of SLNR. It can be seen from the table that when the label noise level in the training set is small (SLNR $\leq$ 20%), standard learning of CNN model parameters achieves higher mAP values compared to label noise robust learning. As an example, when there is no synthetic label noise added to the training set, standard learning leads to more than 3% higher mAP score compared to label noise robust learning. However, as the SLNR value of the training set is higher than a particular value (20%), the considered CNN model with label noise robust learning provides higher multi-label RS image classification accuracies compared to standard learning. For example, when SLNR equals to 80% for the training set, label noise robust learning leads to almost 27% higher mAP value compared to standard learning. These results show that our learning method provides more robust learning of the model parameters for the considered CNN model toward label noise in the training set. Due to the two-stage learning procedure in our method, a

Table 2.5 Multi-label image classification results in terms of mean average precision (mAP (%)) obtained by standard learning and our label noise robust learning for different values of SLNR

SLNR (Training Set)	Standard Learning	Label Noise Robust Learning
0%	99.2	95.6
20%	96.8	89.9
40%	70.9	91.0
60%	66.6	88.3
80%	60.4	87.0

small trustworthy subset of the training set is effectively utilized in its first stage to automatically define noisy labels in the whole training set, which are accurately corrected. Then, employing corrected labels for fine-tuning CNN model parameters on the whole training set leads to leveraging abundant training data without being significantly affected by the label noise.

2.4 Conclusion and Outlook

While OSM provides ample opportunities for use as labels in machine learning-based remote sensing applications, it is necessary to be aware of the challenges the dataset provides. Intrinsic and semi-intrinsic data quality indicators provide insights into the complexity of the OSM mapping process. Meaningful relationships between the indicators and data quality for a test set were derived. The complexity of the interactions did, however, not allow for a reliable prediction of data quality at the level of individual OSM objects. This might change if bigger sample sizes are used. And, while object-level quality prediction requires further research, the developed quality indicators referencing the data region can already support regional quality predictions which are successfully in use in production today.

The proposed deep learning method showed its potential to perform label noise robust multi-label image classification if at least a small set of high-quality labels is available. This shows the potential of the method (i) to overcome the challenges of OSM land-use labels in remote sensing applications and (ii) to provide quality-related feedback for the OSM community. As the OSM community is skeptical toward imports, especially based on automatic labeling, areas flagged as potentially problematic will when presumable be investigated by human mappers and potentially corrected in OSM. Furthermore, these areas can further be analyzed in combination with the intrinsic data quality indicators developed during the project. Approaches described in Chap. 7 might become helpful for this communication. The remote sensing community, on the other hand, can profit from this work through the automated creation of regionalized high-quality image classification models.

Acknowledgments This research was supported by the German Research Foundation DFG within Priority Research Program 1894 *Volunteered Geographic Information: Interpretation, Visualization, and Social Computing* (VGIscience). In this chapter, results from the project *Information Discovery from Big Earth Observation Data Archives by Learning from Volunteered Geographic Information* (IDEAL-VGI, Project number 424966858) are reviewed.

References

Aksoy AK, Ravanbakhsh M, Demir B (2022) Multi-label noise robust collaborative learning for remote sensing image classification. IEEE Trans Neural Netw Learn Syst 1–14. https://doi.org/10.1109/TNNLS.2022.3209992

Audebert N, Le Saux B, Lefèvre S (2017) Joint learning from earth observation and openstreetmap data to get faster better semantic maps. In: IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp 1552–1560. https://doi.org/10.1109/CVPRW.2017.199

Barron C, Neis P, Zipf A (2014) A comprehensive framework for intrinsic openstreetmap quality analysis. Trans GIS 18(6):877–895. https://doi.org/10.1111/TGIS.12073

Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc B (Methodological) 57(1):289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

Brückner J, Schott M, Zipf A, Lautenbach S (2021) Assessing shop completeness in openstreetmap for two federal states in Germany. AGILE: GIScience Series 2:20. https://doi.org/10.5194/agile-giss-2-20-2021

Burgert T, Ravanbakhsh M, Demir B (2022) On the effects of different types of label noise in multi-label remote sensing image classification. IEEE Trans Geosci Remote Sensing 60:1–13. https://doi.org/10.1109/TGRS.2022.3226371

Chen L-C, Zhu Y, Papandreou G, Schroff F, Adam H (2018) Encoder-decoder with atrous separable convolution for semantic image segmentation. In: European Conference on Computer Vision. https://doi.org/10.1007/978-3-030-01234-2_49

Comandur B, Kak AC (2021) Semantic labeling of large-area geographic regions using multiview and multidate satellite images and noisy OSM training labels. IEEE J Sel Topics Appl Earth Obs Remote Sensing 14:4573–4594. https://doi.org/10.1109/JSTARS.2021.3066944

Dong R, Fang W, Fu H, Gan L, Wang J, Gong P (2022) High-resolution land cover mapping through learning with noise correction. IEEE Trans Geosci Remote Sensing 60:1–13. https://doi.org/10.1109/TGRS.2021.3068280

Fonte C, Minghini M, Antoniou V, See L, Patriarca J, Brovelli M, Milcinski G (2016) An automated methodology for converting OSM data into a land use/cover map. In: 6th International Conference on Cartography & GIS, 13–17 June 2016, Albena, Bulgaria

Fonte CC, Bastin L, See L, Foody G, Lupia F (2015) Usability of VGI for validation of land cover maps. Int J Geograph Inform Sci 29(7):1269–1291. ISSN 1365-8816. https://doi.org/10.1080/13658816.2015.1018266

Hollander M, Wolfe DA (1973) Nonparametric statistical methods. Wiley series in probability and mathematical statistics: applied probability and statistics. Wiley, New York. ISBN 0-471-40635-X and 978-0-471-40635-8. https://doi.org/10.1002/9781119196037

Jokar Arsanjani J, Mooney P, Zipf A, Schauss A (2015) Quality assessment of the contributed land use information from openstreetmap versus authoritative datasets. In: Jokar Arsanjani J, Zipf A, Mooney P, Helbich M (eds) OpenStreetMap in GIScience: Experiences, Research, and Applications. Springer, Cham, pp 37–58. https://doi.org/10.1007/978-3-319-14280-73

Kaiser P, Wegner JD, Lucchi A, Jaggi M, Hofmann T, Schindler K (2017) Learning aerial image segmentation from online maps. IEEE Trans Geosci Remote Sensing 55(11):6054–6068. https://doi.org/10.1109/TGRS.2017.2719738

Lee K-H, He X, Zhang L, Yang L (2018) Cleannet: Transfer learning for scalable image classifier training with label noise. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 5447–5456. https://doi.org/10.1109/CVPR.2018.00571

Li H, Dou X, Tao C, Wu Z, Chen J, Peng J, Deng M, Zhao L (2020) RSI-CB: a large-scale remote sensing image classification benchmark using crowdsourced data. Sensors 20(6). https://doi.org/10.3390/s20061594

Lin J, Yu T, Wang ZJ (2022) Rethinking crowdsourcing annotation: partial annotation with salient labels for multilabel aerial image classification. IEEE Trans Geosci Remote Sensing 60:1–12. https://doi.org/10.1109/TGRS.2022.3191735

Liu S, Niles-Weed J, Razavian N, Fernandez-Granda C (2020) Early-learning regularization prevents memorization of noisy labels. In: International Conference on Neural Information Processing Systems. https://doi.org/10.48550/arXiv.2007.00151

Neis P, Zielstra D, Zipf A (2013) Comparison of volunteered geographic information data contributions and community development for selected world regions. Fut Int 5(2):282–300. https://doi.org/10.3390/fi5020282

OpenStreetMap Contributors (2022a) List of OSM-based services. https://wiki.openstreetmap.org/wiki/List_of_OSM-based_services

OpenStreetMap Contributors (2022b) Category:OSM processing. https://wiki.openstreetmap.org/wiki/Category:OSM_processing

Ridnik T, Ben-Baruch E, Zamir N, Noy A, Friedman I, Protter M, Zelnik-Manor L (2021) Asymmetric loss for multi-label classification. In: IEEE/CVF International Conference on Computer Vision, pp 82–91. https://doi.org/10.1109/ICCV48922.2021.00015

Schott M, Grinberger AY, Lautenbach S, Zipf A (2021) The impact of community happenings in OpenStreetMap—establishing a framework for online community member activity analyses. ISPRS Int J Geo-Inform 10(3):164. https://doi.org/10.3390/ijgi10030164

Schott M, Lautenbach S, Größchen L, Zipf A (2022) Openstreetmap element vectorisation—a tool for high resolution data insights and its usability in the land-use and land-cover domain. Int Arch Photogramm Remote Sensing Spatial Inform Sci 48:4. https://doi.org/10.5194/isprs-archives-XLVIII-4-W1-2022-395-2022

Senaratne H, Mobasheri A, Ali AL, Capineri C, Haklay MM (2017) A review of volunteered geographic information quality assessment methods. Int J Geograph Inform Sci 31(1):139–167. https://doi.org/10.1080/13658816.2016.1189556

Wan T, Lu H, Lu Q, Luo N (2017) Classification of high-resolution remote-sensing image using openstreetmap information. IEEE Geosci Remote Sensing Lett 14 (12):2305–2309. https://doi.org/10.1109/LGRS.2017.2762466

Wei H, Feng L, Chen X, An B (2020) Combating noisy labels by agreement: a joint training method with co-regularization. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 13723–13732. https://doi.org/10.1109/CVPR42600.2020.01374

Yao J, Wang J, Tsang IW, Zhang Y, Sun J, Zhang C, Zhang R (2019) Deep learning from noisy image labels with quality embedding. IEEE Trans Image Process 28(4):1909–1922. https://doi.org/10.1109/TIP.2018.2877939

Zhang R, Chen Z, Zhang S, Song F, Zhang G, Zhou Q, Lei T (2020a) Remote sensing image scene classification with noisy label distillation. Remote Sensing 12(15). https://doi.org/10.3390/rs12152376

Zhang X, Wei Y, Yang Y, Wu F (2020b) Rethinking localization map: towards accurate object perception with self-enhancement maps. arXiv preprint arXiv:2006.05220. https://doi.org/10.48550/arXiv.2006.05220

Zhou B, Khosla A, Lapedriza A, Oliva A, Torralba A (2016) Learning deep features for discriminative localization. In: IEEE Conference on Computer Vision and Pattern Recognition, pp 2921–2929. https://doi.org/10.1109/CVPR.2016.319