EFFECT OF THE TRAINING SET CONFIGURATION ON SENTINEL-2-BASED URBAN LOCAL CLIMATE ZONE CLASSIFICATION

C. P. Qiu[a], M. Schmitt[a], P. Ghamisi[b], X. X. Zhu[a,b]

[a] Signal Processing in Earth Observation, Technical University of Munich (TUM), Arcisstr. 21, 80333 Munich, Germany - (chunping.qiu, m.schmitt)@tum.de
[b] Remote Sensing Technology Institute (IMF), German Aerospace Center (DLR), Oberpfaffenhofen, 82234 Wessling, Germany - xiao.zhu@dlr.de

Commission II, WG II/6

KEY WORDS: Label Noise, Canonical Correlation Forests (CCFs), Local Climate Zones (LCZs), Sentinel-2, Spectral features, Classification, Open Street Map (OSM)

ABSTRACT:

As any supervised classification procedure, also Local Climate Zone (LCZ) mapping requires reliable reference data. These are usually created manually and inevitably include label noise, caused by the complexity of the LCZ class scheme as well as variations in cultural and physical environmental factors. This study aims at evaluating the impact of the training set configuration, i.e. training sample number and imbalance, on the performance of Canonical Correlation Forests (CCFs) for a classification of the 11 urban LCZ classes. Experiments are carried out based on globally available Sentinel-2 imagery. Besides multi-spectral observations, different index measures extracted from the images as well as the Global Urban Footprint (GUF) and Open Street Map (OSM) layers are fed into the CCFs classifier. The results show that different LCZs favor different configurations in terms of training sample number and balance. Based on the findings, majority voting of different predictions from different configurations is proposed and performed. This way, a significant accuracy improvement can be achieved.

1. INTRODUCTION

Local Climate Zone (LCZ) mapping (Stewart and Oke, 2012), originally developed for meta-data communication of observational Urban Heat Island (UHI) studies, has gained great interest in the field of remote sensing. The 17 LCZ classes, as displayed in Fig. 1, are based on climate-relevant surface properties mainly related to 3D surface structure (e.g. height and density of buildings and trees) as well as surface cover (e.g., pervious or impervious). Recently, researchers have started to use the LCZ scheme to classify the internal structure of urban areas, providing critical support for various applications such as urban climatology, infrastructure planning, disaster mitigation, etc.

Supervised classification methods using remote sensing data as input provide a valuable tool for LCZ mapping and have been widely studied in the 2017 Data Fusion Contest organized by the Image Analysis and Data Fusion Technical Committee (IADFTC) of the IEEE Geoscience and Remote Sensing Society (GRSS) (Yokoya et al., 2017). The supervised strategy requires a training dataset in order to train a classifier, which can later be used to predict the labels of unseen samples. Each sample in the training dataset is defined by a feature vector and its class label. This training dataset is crucial for the classification accuracy as well as the generalization ability of the trained classifier. In order to guarantee a satisfying LCZ mapping accuracy, the training data should be of sufficient size and provide well balanced sample numbers for all 17 LCZ classes.Unlike training data for other land cover/land use classifications, LCZ training samples are not easy to extract from existing databases. Therefore, it is common to generate LCZ reference data manually (Bechtel et al., 2015), in spite of being time consuming. Unfortunately, obtaining reference data of high quality is challenging, especially when it comes to global scale where variations in cultural and physical environmental factors exist.

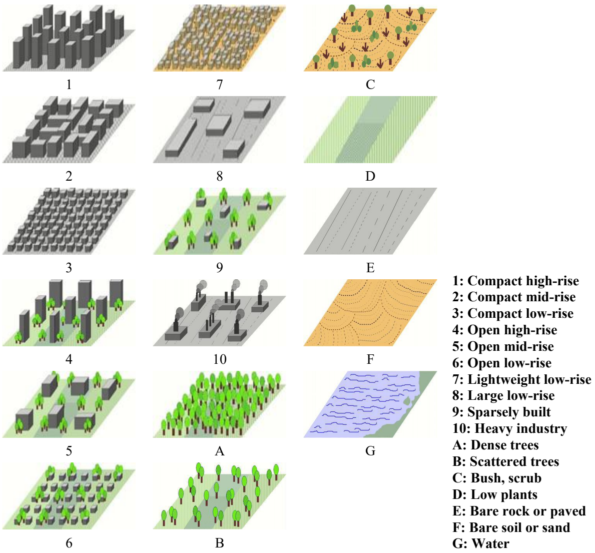


Figure 1. Visualization of the LCZ concept (Stewart, 2011).

Few works have systematically analyzed the effect of the training set configuration on classification, such as training sample number and balance among different classes, and much less are specifically dealing with LCZ mapping. Existing literature either deals with the problem from a theoretical perspective (Natarajan et al., 2013), or focuses on the relation between label noise and classification complexity using some benchmark dataset (Garcia et al., 2015). In the remote sensing community, two closely re-

This contribution has been peer-reviewed.
https://doi.org/10.5194/isprs-archives-XLII-2-931-2018 | © Authors 2018. CC BY 4.0 License.

lated works are (Goldblatt et al., 2016) and (Pelletier et al., 2017). The former shows that, for built-up area detection, increasing the number of training samples of one class will improve the classifier's performance even though it will introduce imbalance between classes. The latter one investigates the effect of training class label noise for land cover mapping, with a well balanced training dataset. While inspiring, these works provide limited guidance for the training set configuration necessary for successful LCZ mapping.

Aiming for global LCZ mapping for which training data is costly and resource intensive to collect, our work intends to provide an answer to these questions: How do the training set size and the distribution of training samples across the classes impact the LCZ classification performances? For simplicity, we focus on the first ten LCZ classes, which are referred to as urban LCZ classes in this paper. In addition, we add two background classes, namely vegetation and water, to achieve land cover completeness. As a framework for our investigations, we use Canonical correlation Forests (CCFs) (Rainforth and Wood, 2015). We feed different CCF classifiers with different training set configurations using the globally available imagery provided by the Sentinel-2 mission (Radoux et al., 2016). The results achieved in this paper also provide perspectives for other large scale classification tasks employing different classifiers.

2. DATASET AND CLASSIFICATION FRAMEWORK

In this section, we describe the test dataset and the classification framework we use to carry out our experiments.

2.1 Study Areas and LCZ Dataset

Our study areas are spread over 7 cities located in the heart of Europe. They are depicted in Fig. 2. For all those cities, we have downloaded Sentinel-2 imagery from ESA's SciHub (`https://scihub.copernicus.eu/`). In addition, we were allowed to access DLR's Global Urban Footprint (GUF), a binary layer derived from TanDEM-X data, which indicates urban areas (Klotz et al., 2016). Finally, we have downloaded the Open Street Map layers *building* and *land-use* from OpenStreetMap Data Extracts (`http://download.geofabrik.de/`) for each city.

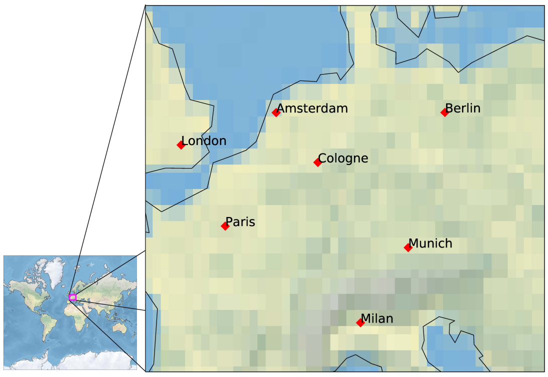


Figure 2. The seven test cities distributed across Europe.

The LCZ ground truth labels available for selected neighborhoods in the 7 cities are taken from the LCZ42 dataset (Zhu, 2018). The number of samples available for each class over the seven cities can be seen in Fig. 3, in which the vegetation class combines the LCZ classes A, B, C, D, and F. Figure 3 illustrates the variability of both the sample number and the class distribution among different cities, which highlights the importance of assessing the effect of training set configurations. It should be noted that in these 7 cities, LCZ class 7 (lightweight low-rise), which mostly indicates slums, does not exist.

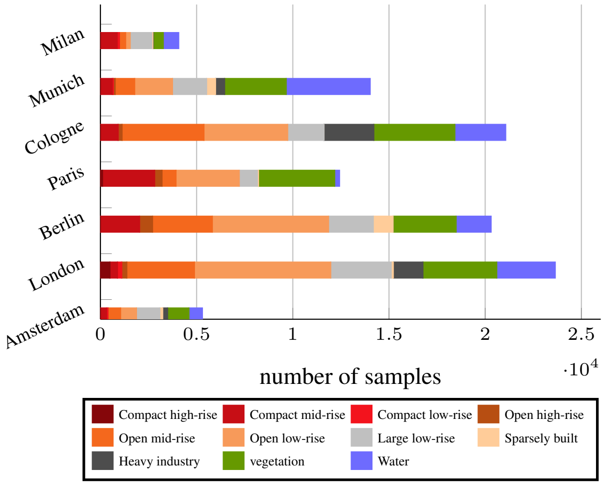


Figure 3. Urban LCZ training sample number of seven cities.

2.2 Classification Framework

We use a Canonical Correlation Forests (CCF) classifier (Rainforth and Wood, 2015) as framework for our investigations, since it has shown strong potential in the IEEE-GRSS Data Fusion contest (Yokoya et al., 2017). As the most important hyperparameter, we fix the number of trees to 20. Furthermore, we define 18 features as input to the classifier: Input Features and datasets are described as follows.

1. *Spectral reflectance*
   10 bands of Sentinel-2 imagery are used in this study: B2, B3, B4 and B8 with 10 m Ground Sampling Distance (GSD) and B5, B6, B7, B8a, B11 and B12 with 20 m GSD. The 20 m bands are up-sampled to 10 m GSD. The bands B1, B9 and B10 are not considered in this study because they contain mostly information about the atmosphere and thus bear little relevance to LCZ classification.

2. *Indices*
   The well-established indices Normalized Difference Vegetation Index (NDVI), Enhanced Vegetation Index (EVI), Modified Normalized Difference Water Index (MNDWI), Normalized Difference Built Index (NDBI) and Bare-Soil Index (BSI) are also considered (Tucker, 1979), since they can provide indications about vegetation, water, buildings and soil, respectively (Yokoya et al., 2017, Bechtel et al., 2015).

3. *Other auxiliary data*
   Besides the Open Street Map layers *buildings* and *land use*, we also use the Global Urban Footprint (GUF) data (Klotz et al., 2016) as an additional input feature. Both OSM and GUF are re-sampled to 10 m GSD.

3. TRAINING SET CONFIGURATION ANALYSIS

In order to evaluate the effect of training sample number and imbalance on the performance of the CCFs for urban LCZ classification, different configurations of training samples are designed

Table 1. Setups for training sample number and balance effect analysis. For setup I and II, the experiment was carried out five times. Each time, the samples used are randomly chosen from the original dataset. For setup III, the experiment was carried out ten times. Each time, $\frac{1}{10}$ of each class sample was randomly chosen and used.

Configuration name	Sample Number	Imbalance-Correction	Description
I	$n_{min} \times 11$	Yes	n_{min} is the minimum of the sample number of all classes.
II	$\sum_{i=1}^{11} min\{n_i, n_{median}\}$	Partly	n_{median} is the median of the sample number of all classes. n_i is the training sample number of each class, $i \in \{1, 2, 3, ...11\}$.
III	$\sum_{i=1}^{11}(n_i \times \frac{1}{10})$	No	
baseline	N	No	N is the total number of the original training samples.

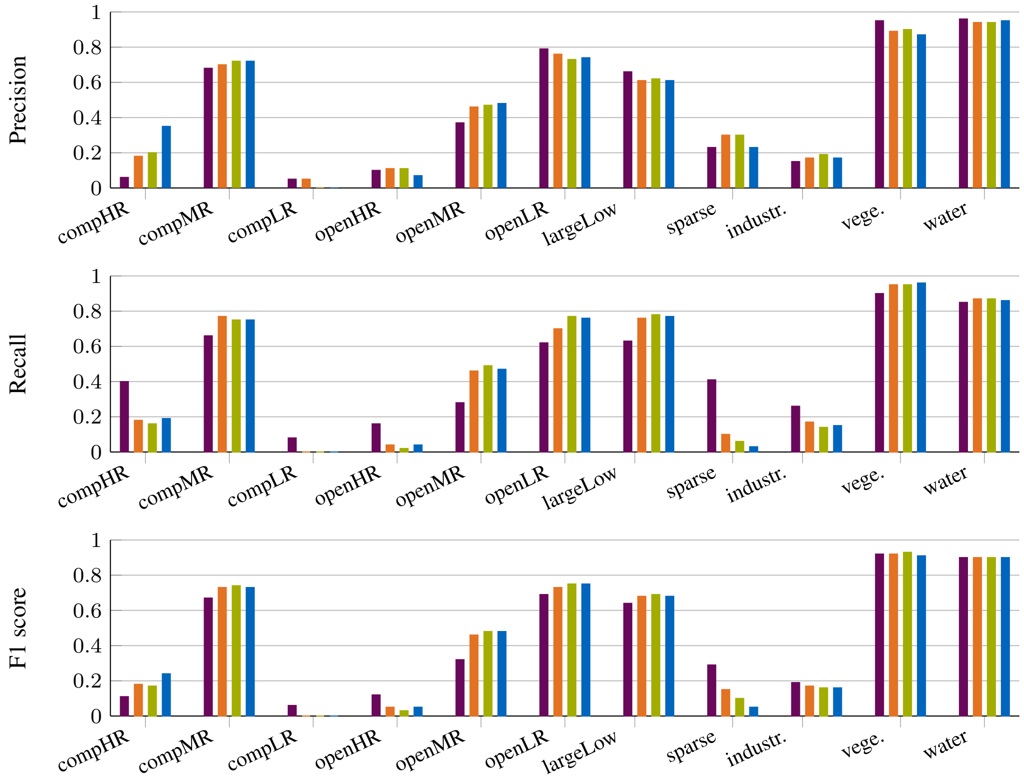


Figure 4. Precision, recall and F1 score of different classes resulted from different training set. For each class, from left to right, corresponding to configuration I, II, III and baseline in Table. 1. These values are averaged over seven cities.

based on different number of samples and different amounts of balancing applied to the different classes. These configurations are summarized in Table. 1.

By comparing the classification accuracies of configurations I, II and III with a baseline configuration in Table. 1, the relative importance of training sample balance and number can be studied. For all experiments, we rely on cross validation, i.e., each time samples from six cities are used for training while those from the seventh city are used for testing. The difference of classification accuracy corresponding to different configurations can be seen in Fig. 4.

Comparing the green and blue bars in Fig. 4, it can be seen that no big difference exists between configuration III and the baseline. This indicates that training set size plays a limited role under the stated experimental conditions. From Fig. 4, it can also be seen that different classes favor different configurations. Specifically, based on the standard of recall and F1 score, configuration I outperforms the other configurations for compact low rise, open high rise, sparsely built and industry, while it provides slightly worse results for the other classes. This means that training sample balance is important for classes with few samples.

Based on these results, it can be concluded that given a certain amount of unbalanced training samples for tasks like LCZ classification, a potential classification improvement can be achieved by elaborately preparing the training dataset, instead of directly using all the training samples when training the classifiers.

https://doi.org/10.5194/isprs-archives-XLII-2-931-2018 |

Table 2. Classification accuracy after major voting and the increased ratio compared to the baseline accuracy.

city	after majority voting			increased ratio		
	OA	Kappa	AA	OA	Kappa	AA
Amsterdam	0.70	0.65	0.47	2.61	3.41	1.96
London	0.72	0.66	0.48	7.35	9.54	3.72
Berlin	0.64	0.56	0.49	2.54	3.66	4.52
Paris	0.82	0.77	0.56	1.33	1.76	2.27
Cologne	0.68	0.62	0.60	3.25	4.32	2.93
Munich	0.83	0.79	0.53	3.96	5.26	7.29
Milan	0.84	0.80	0.65	2.15	2.76	7.10
Mean	**0.75**	**0.69**	**0.54**	**3.31**	**4.39**	**4.26**

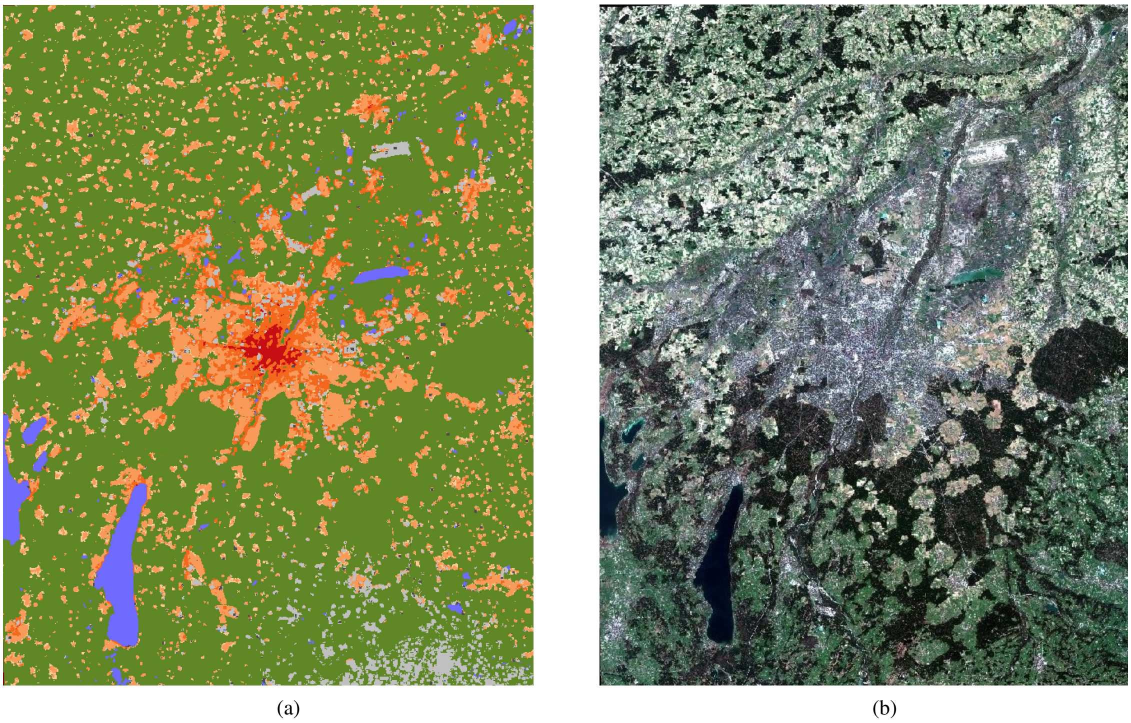


Figure 5. The LCZ map of Munich (a) produced using majority voting, and the corresponding Sentinel-2 RGB imagery (b). The legend is the same with that in Fig. 3

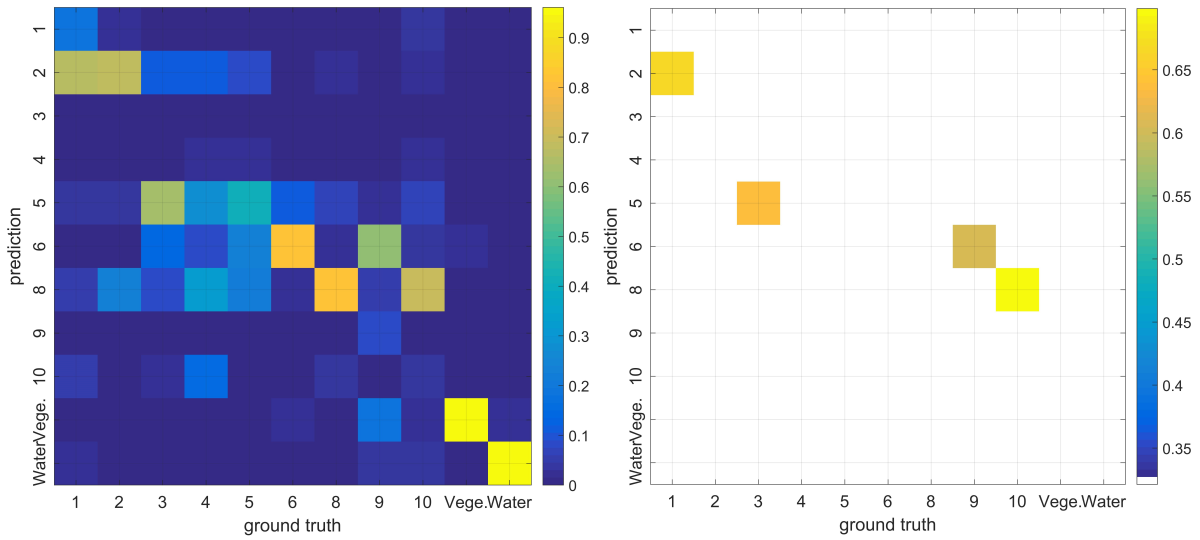
4. IMPROVING URBAN LCZ CLASSIFICATION

Different CCFs can be trained using the different configurations from Table. 1. A final classification can be achieved by applying majority voting on the results of all the 20 configurations. The classification accuracy after majority voting and its improvement relative to the baseline accuracy is shown in Table. 2. Using this majority voting-based result, an LCZ classification map over the city of Munich, Germany, is shown in Fig. 5.

5. DISCUSSION

The final result achieved in Section 4 shows that a promising accuracy can be achieved by majority voting of CCFs trained with different training set configurations. As the values in Table. 2 show, about 4% of improvement of Overall Accuracy (OA), Kappa coefficient and Averaged Accuracy (AA) can be achieved in this case. This improvement is resulted from the increased accuracy of classes with fewer samples, such as compact high rise and compact low rise, as Section 3 shows training set balance plays a crucial role to the classification accuracy for classes with fewer samples.

Nevertheless, even after majority voting, averaged accuracy are still less than 50% for about 3 of the 7 test cities. The misclassification between classes can be analyzed using confusion matrices of the classification results. For conciseness Figure 6(a) depicts the combined confusion matrix of all the 7 test cases, and Figure 6(b) highlights the misclassification errors higher than 30%. Class 1 (compact high rise) is falsely classified into class 2 (compact middle rise). This is resulted from the challenge of distinguishing height difference using Sentinel-2 images, since high rise and middle rise are quite similar in the two dimensional optical images. Besides, class 9 (sparsely built) and 10 (heavy industry) are falsely classified into class 6 (open low rise) and 8 (large low rise), respectively. This is due to inter-class similarity, as they appear quite similar, as can be seen in Fig. 1. Besides, class 3 (compact low rise) is unexpectedly misclassified into class 5 (open middle rise).The first reason may be also related to the difficulty to distinguish low rise and middle rise. Another possible reason is that the sample number of class 3 is the fewest and much



(a)

(b)

Figure 6. Combined confusion matrix of 7 cities (a) and the cases with a misclassification error higher than 30% (b).

fewer than the others (only 466 samples for all 7 cities), as can also be seen in Fig. 3. As a result, the intra-class variability is not well learned during training.

To solve these problems, one possible solution is to include additional datasets such as Synthetic Aperture Radar images to make use of radar’s unique range measurements. Another solution is to adapt the LCZ scheme considering the feasibility of optical images, or a multi-level classification might be beneficial. Last but not least, negative human influence on ground truth should be weaken to guarantee the quality of the training samples across cities (Bechtel et al., 2017).

6. SUMMARY AND CONCLUSION

This paper presents an investigation of the effect of the training set configuration on urban LCZ classification. It intends to provide potential answers to the question: how do the training set size and potential class imbalances affect the classification results? Based on the powerful classification framework Canonical Correlation Forests, a series of experiments has been carried out over 7 cities in Europe. Based on the experimental results, it can be concluded that the training set size is not as important as the training sample balance, especially for classes with relatively few samples. Based on the findings of this analysis, majority voting of classification results from different configurations was investigated as well, which lead to a significant improvement in the achievable classification accuracy. Still, even majority voting does not provide the perfect solution to urban LCZ classification, which motivates investigations towards more advanced classifiers and the fusion of complementary data in the future.

ACKNOWLEDGEMENTS

This work is jointly supported by the China Scholarship Council, the Helmholtz Association under the framework of the Young Investigators Group SiPEO (VH-NG-1018, www.sipeo.bgu.tum.de), and the European Research Council (ERC) under the European Union’s Horizon 2020 research and innovation programme (grant agreement No. ERC-2016-StG-714087, Acronym: *So2Sat*).

REFERENCES

Bechtel, B., Alexander, P. J., Böhner, J., Ching, J., Conrad, O., Feddema, J., Mills, G., See, L. and Stewart, I., 2015. Mapping local climate zones for a worldwide database of the form and function of cities. *ISPRS International Journal of Geo-Information* 4(1), pp. 199–219.

Bechtel, B., Demuzere, M., Sismanidis, P., Fenner, D., Brousse, O., Beck, C., Van Coillie, F., Conrad, O., Keramitsoglou, I., Middel, A. et al., 2017. Quality of crowdsourced data on urban morphologythe human influence experiment (huminex). *Urban Science* 1(2), pp. 15.

Garcia, L. P., de Carvalho, A. C. and Lorena, A. C., 2015. Effect of label noise in the complexity of classification problems. *Neurocomputing* 160, pp. 108–119.

Goldblatt, R., You, W., Hanson, G. and Khandelwal, A. K., 2016. Detecting the boundaries of urban areas in india: A dataset for pixel-based image classification in google earth engine. *Remote Sensing* 8(8), pp. 634.

Klotz, M., Kemper, T., Geiß, C., Esch, T. and Taubenböck, H., 2016. How good is the map? a multi-scale cross-comparison framework for global settlement layers: Evidence from central europe. *Remote Sensing of Environment* 178, pp. 191–212.

Natarajan, N., Dhillon, I. S., Ravikumar, P. K. and Tewari, A., 2013. Learning with noisy labels. In: *Advances in neural information processing systems*, pp. 1196–1204.

Pelletier, C., Valero, S., Inglada, J., Champion, N., Marais Sicre, C. and Dedieu, G., 2017. Effect of training class label noise on classification performances for land cover mapping with satellite image time series. *Remote Sensing* 9(2), pp. 173.

Radoux, J., Chomé, G., Jacques, D. C., Waldner, F., Bellemans, N., Matton, N., Lamarche, C., D’Andrimont, R. and Defourny, P., 2016. Sentinel-2’s potential for sub-pixel landscape feature detection. *Remote Sensing*.

Rainforth, T. and Wood, F., 2015. Canonical correlation forests. *arXiv preprint arXiv:1507.05444*.

Stewart, I. D., 2011. Local climate zones: Origins, development, and application to urban heat island studies. *Paper presented at the Annual Meeting of the American Association of Geographers*.

Stewart, I. D. and Oke, T. R., 2012. Local climate zones for urban temperature studies. *Bulletin of the American Meteorological Society* 93(12), pp. 1879–1900.

Tucker, C. J., 1979. Red and photographic infrared linear combinations for monitoring vegetation. *Remote sensing of Environment* 8(2), pp. 127–150.

Yokoya, N., Ghamisi, P. and Xia, J., 2017. Multimodal, multitemporal, and multisource global data fusion for local climate zones classification based on ensemble learning. *Proc. 37th annual symposium of the IEEE Geoscience and Remote Sensing Society (GRSS), Fort Worth, Texas, USA, 23-28 July.*

Zhu, X. X., 2018. So2sat LCZ42 dataset, to appear.