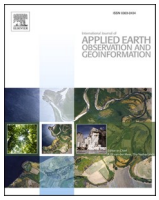
International Journal of Applied Earth Observations and Geoinformation 103 (2021) 102482

Contents lists available at ScienceDirect

International Journal of Applied Earth Observations and Geoinformation

journal homepage: www.elsevier.com/locate/jag



Application of training data affects success in broad-scale local climate zone mapping

Chunxue Xu [a,*], Perry Hystad [b], Rui Chen [c], Jamon Van Den Hoek [a], Rebecca A. Hutchinson [d,e], Steve Hankey [f], Robert Kennedy [a]

[a] College of Earth, Ocean, and Atmospheric Sciences, Oregon State University, Corvallis, OR, United States
[b] College of Public Health and Human Sciences, Oregon State University, Corvallis, OR, United States
[c] Department of Computer Science, Tufts University, Medford, MA, United States
[d] School of Electrical Engineering and Computer Science, Oregon State University, Corvallis, OR, United States
[e] Department of Fisheries, Wildlife, and Conservation Sciences, Oregon State University, Corvallis, OR, United States
[f] School of Public and International Affairs, VA Tech, VA, United States

ARTICLE INFO

Keywords:
Local climate zone
Machine learning
Training areas
Crowdsourced data
Spatial autocorrelation

ABSTRACT

Satellite imagery has been widely used to map urbanization processes. To address the urgent need for urban landscape mapping that goes beyond urban footprint analysis, the local climate zone (LCZ) scheme has been increasingly used to reveal the urban forms and functions important to urban heat islands and micro-climates across the globe. As with most supervised classification strategies, proper application of training data is critical for the success of LCZ classification models. However, the collection and application of LCZ training areas brings with it two challenges that may affect mapping success. First, because digitizing training areas is a time-consuming task, there is a broad effort in the LCZ mapping community to create a crowdsourced data collection among different experts. However, this strategy likely leads to inconsistencies in labels that could weaken models. Second, the LCZ labeling process typically involves the delineation of large zones from which multiple training samples are drawn, but those samples are likely spatially autocorrelated and lead to overly optimistic estimates of model accuracy. Although both effects – inconsistent labeling and spatial autocorrelation – are theoretically possible, it is unknown whether they substantially affect accuracy. We investigated both issues, specifically asking: (i) how do the discrepancies of LCZ labeling by different experts impact broad-scale LCZ mapping? (ii) to what extent does spatial correlation affect model prediction power? We used two classifiers (Random Forests and ResNets) to map eight metropolitan areas in the US into LCZs, comparing training areas drawn by different or consistent interpreters, and data splitting strategy using rules that allow or reduce spatial autocorrelation. We found large discrepancies among results built from crowdsourced training areas digitized by different experts; improving the consistency of labels can lead to substantial improvements in LCZ classification accuracy. Second, we found that spatial autocorrelation can boost the apparent accuracy of the classifier by 16% to 21%, leading to erroneous interpretation of mapping results. The two effects interplay as well: spatial autocorrelation in the raw data can lead to an underestimation of the model's predictive error when modeling with crowdsourced training areas of high inconsistency. Due to the uncertainty in the labeling process and spatial autocorrelation in derived training data, broad-scale LCZ mapping results should be interpreted with caution.

1. Introduction

More than half of the world's population lives in urban areas, and 2.5 billion new urban dwellers are projected by 2050 (United Nations, 2014). The shifts in population and land use associated with this unprecedented urbanization are fundamentally transforming relationships between cities and the global environment (Seto et al., 2010). Urbanization has brought formidable challenges in climate change, energy consumption, environmental pollution, and social problems such as justice and human health (Agathangelidis et al., 2019; Forman, 2014). Artificial surfaces in urban areas create distinctive local climates, and there has been increasing attention in urban climate science on revealing

* Corresponding author.
E-mail address: xuch@oregonstate.edu (C. Xu).

https://doi.org/10.1016/j.jag.2021.102482
Received 24 March 2021; Received in revised form 8 June 2021; Accepted 6 August 2021
Available online 29 August 2021

the characteristics and impacts of urban structure, form, and functions on climate patterns and processes (Dhar and Khirfan, 2017; Middel et al., 2018).

Earth Observation (EO) data have been increasingly used to reveal urban form at the regional, country, and global levels. Numerous urban land cover products have been developed and applied to urban climate studies. For example, The German Aerospace Center developed the Global Urban Footprint (GUF) raster dataset and revealed global human settlements at a spatial resolution of 0.4° (~12 m) (Esch et al., 2017). As an extension of GUF, the World Settlement Footprint 2015 was developed by Marconcini et al. (2020) as a global map of human settlements for the year 2015 using open-and-free optical and radar satellite imagery at 10 m resolution. In these and other broad-scale remotely sensed maps, the urban extent is often presented as a binary built-up/non-built-up product that defines areas within and outside the "urban footprint." Such binary urban footprint labels provide an inventory of human settlement locations and total extents at a given time, but fail to capture the internal structure and texture of cities worldwide.

To address the urgent need for a comprehensive worldwide urban product with more detailed information of urban structure than a simple urban footprint map, the World Urban Database and Access Portal Tools project (WUDAPT, http://www.wudapt.org/) has been developed to provide consistent knowledge of urban internal layout for cities across the globe (Bechtel et al., 2015). The description of the urban layout in WUDAPT is based on the local climate zone (LCZ) scheme (Fig. 1), which was initially developed by Stewart and Oke (2012) to address the inadequacies of a binary urban/non-urban description in urban temperature studies. The LCZ scheme divides the urban landscape into 17 homogeneous types based on urban layout, land cover, surface materials, and human activities. It seeks to provide a generic and culturally-neutral classification of the urban landscape for climate studies, which means that LCZ classes can be applied universally instead of being

Built Types	Nadir View	Physical Properties	Land Cover Types	Nadir View	Physical Properties
LCZ1 Compact high-rise		a: 0.2-0.4 b: >2 c: 40-60 d: 40-60 e: <10 f: >25 g: 8	LCZA Dense trees		a: <0.4 b: >1 c: <10 d: <10 e: >90 f: 3-30 g: 8
LCZ2 Compact midrise		a: 0.3-0.6 b: 0.75-2 c: 40-70 d: 30-50 e: <20 f: 10-25 g: 6-7	LCZB Scattered trees		a: 0.5-0.8 b: 0.25-0.75 c: <10 d: <10 e: >90 f: 3-15 g: 5-6
LCZ3 Compact low-rise		a: 0.2-0.6 b: 0.75-1.5 c: 40-70 d: 20-50 e: <30 f: 3-10 g: 6	LCZC Bush, scrub		a: 0.7-0.9 b: 0.25-1 c: <10 d: <10 e: >90 f: >2 g: 4-5
LCZ4 Open high-rise		a: 0.5-0.7 b: 0.75-1.25 c: 20-40 d: 30-40 e: 30-40 f: >25 g: 7-8	LCZD Low plants		a: >0.9 b: <0.1 c: <10 d: <10 e: >90 f: <1 g: 3-4
LCZ5 Open midrise		a: 0.5-0.8 b: 0.3-0.75 c: 20-40 d: 30-50 e: 20-40 f: 10-25 g: 5-6	LCZE Bare rock or paved		a: >0.9 b: <0.1 c: <10 d: >90 e: <10 f: <0.25 g: 1-2
LCZ6 Open low-rise		a: 0.6-0.9 b: 0.3-0.75 c: 20-40 d: 30-50 e: 20-40 f: 3-10 g: 5-6	LCZF Bare soil or sand		a: >0.9 b: <0.1 c: <10 d: <10 e: >90 f: >0.25 g: 1-2
LCZ7 Lightweight low-rise	—	a: 0.2-0.5 b: 1-2 c: 60-90 d: <20 e: <30 f: 2-4 g: 4-5	LCZG Water		a: >0.9 b: <0.1 c: <10 d: <10 e: >90 f: - g: 1
LCZ8 Large low-rise		a: >0.7 b: 0.1-0.3 c: 30-50 d: 40-50 e: <20 f: 3-10 g: 5-6			
LCZ9 Sparsely built		a: >0.8 b: 0.1-0.25 c: 10-20 d: <20 e: 60-80 f: 3-10 g: 5-6			
LCZ10 Heavy industry		a: 0.6-0.9 b: 0.2-0.5 c: 20-30 d: 20-40 e: 40-50 f: 5-15 g: 5-6			

a: Sky view factor
b: Aspect ratio
c: Building surface fraction
d: Impervious surface fraction
e: Pervious surface fraction
f: Height of roughness elements
g: Terrain roughness class

Fig. 1. Local climate zone scheme (modified after Stewart and Oke, 2012). Examples of Google Earth satellite imagery are shown for LCZ classes, representing the corresponding zones for the eight metropolitan areas in the U.S. that make up the study area. The nadir-view image for LCZ7 is not shown as it is rare in the study area.

culture- or region-specific (Stewart and Oke, 2012). In 2015, WUDAPT proposed the initial phase of the LCZ classification workflow using satellite imagery, including the preparation of training areas and Landsat 8 imagery, classification using Random Forests, and post-classification processing. The pipeline has been widely used for LCZ mapping by remote sensing scholars in diverse urban regions (Ren et al., 2019; Rosentreter et al., 2020; Yoo et al., 2019). Although there has been increasing attention toward LCZ classification, most research has focused on investigating classifiers and methods, whereas limited attention has been placed on the application of training data.

High-quality training areas are the basis for generating training data for LCZ mapping in a typical WUDAPT workflow. Training areas are polygons manually digitized by experts from very-high-resolution imagery to represent LCZ classes, which are then used as reference geometries in sampling data from satellite imagery for model training and evaluation. The process of identifying training areas is time-consuming and requires that experts have knowledge specific to the LCZ scheme and cities of interest. To decentralize training area collection, WUDAPT provides a crowdsourcing platform for urban experts to share and access training area sets, thereby incorporating expert knowledge from local communities in different cities around the world (Bechtel et al., 2017). With increasing attention toward broad-scale LCZ mapping and modeling for multiple cities within (Demuzere et al., 2019, 2020a) or across different continents (Rosentreter et al., 2020; Yoo et al., 2019), WUDAPT is an essential big data source of labels for global-scale LCZ mapping. As of writing, there are training areas available for more than 120 cities worldwide on WUDAPT.

A central challenge in broad-scale LCZ mapping compared to mapping an individual city is maintaining thematic consistency across training areas. Ideally, LCZ mapping requires reference polygons to be digitized and labeled in a consistent manner to minimize the impacts of spatial variation for a certain LCZ class across different cities. Since urban areas are highly heterogeneous and not readily labeled with LCZ classes using satellite imagery, expert knowledge is required in labeling various zones within a city. However, the bias in human interpretation of satellite imagery is inevitable in the labeling process (Zhu et al., 2020). Large discrepancies have been found in LCZ maps using training areas created by different experts (Bechtel et al., 2017), and WUDAPT has emphasized that training areas are available on an *as-is* basis and that the quality cannot be guaranteed. With the goal of developing globally-consistent and comparable LCZ maps, we must better understand how discrepancies among training areas generated from different experts can impact LCZ mapping results and applications worldwide.

Another related challenge associated with training areas in broad-scale LCZ mapping is that spatial autocorrelation in training data can potentially lead to an overestimation of model performance. Most LCZ mapping studies use machine learning pipelines for classification, and model evaluation is typically performed by randomly splitting the observations into distinct training and testing sets. To provide an unbiased evaluation of the model, the testing set should be independent of the training set, but have the same probability distribution (Xu and Goodacre, 2018). However, spatial data often show internal dependence structures (nearby observations have similar characteristics than distant observations). Spatial autocorrelation can potentially violate the critical assumption of independence in the modeling process, and lead to overestimated model performances if different dependency structures present in the training process and the prediction set-up. In LCZ mapping studies, there have been two widely used strategies for generating training and testing subsamples. One strategy is to collect all observations within training areas to create a sample pool, and then divide the sample pool into training and testing sets (e.g., Demuzere et al., 2019). The other strategy is polygon held-out, splitting the polygons to reduce spatial dependence among samples for training and testing purposes (e.g., Yoo et al., 2019). Because samples drawn within the same polygon are close to each other, the chance of spatial autocorrelation among samples is higher for the first strategy. Spatial autocorrelation could potentially inflate apparent model performance, but it is not clear whether this impact is consequential in broad-scale LCZ mapping.

Here, we report on a study to examine the above challenges with training area labeling and application in broad-scale LCZ mapping. For eight major metropolitan areas in the U.S., we mapped LCZs using Landsat imagery and explored model performance using two different types of labelers (i.e., crowdsourced training areas digitized by different urban experts, referred to as crowdsourced TAs; and refined training areas interpreted by one expert, referred to as refined TAs), and splitting training areas in two different ways to build training and testing sets (i.e., splitting the sample pool and splitting the polygon pool). For each of these four combinations, we used both a pixel-based classification algorithm (i.e., Random Forests) and a scene classification algorithm (i.e., residual neural networks) for LCZ classification. The former serves as a standard pixel-based classifier using spectral features as input, whereas the latter is a deep learning algorithm that can incorporate high-level semantic features such as neighborhood information in image classification. We aimed to address the following research questions: (i) how do the discrepancies of LCZ labeling by different experts affect the accuracy of broad-scale LCZ mapping?, and (ii) to what extent does spatial autocorrelation impact the model prediction power in broad-scale LCZ mapping studies?

2. Data and methods

2.1. Study area

We chose eight metropolitan statistical areas defined in March 2020 by the US Office of Management and Budget (Fig. 2) as study sites. Each metropolitan area is a geographic region with a high-density population nucleus at its core and adjacent socioeconomically connected communities. Our sites represent a range of environmental features such as temperature, precipitation, and vegetation.

2.2. Data

2.2.1. Training areas

Two different training area sets were used in this study. One dataset was a crowdsourced dataset created by different urban experts (available on WUDAPT) and workers on the Amazon Mechanical Turk (MTurk), which was used in support of a CONUS-wide mapping project (Demuzere et al., 2020a). We used crowdsourced TAs available for the study area in this dataset (Demuzere et al., 2020b), from which expert labels are available in six urban areas: New York, Chicago, Houston, Washington DC, Philadelphia, and Los Angeles. By examining the crowdsourced TAs, we found two problems that might impact data quality: (1) some training areas with similar surface properties are labeled inconsistently by different urban experts; (2) some training areas are mixed zones with various LCZ types inside the polygons (Fig. 3).

This first dataset was then curated and augmented to produce the second dataset–refined TAs. For six of the metropolitan areas where expert labels from the first dataset were available, one expert on our team used Google Earth Pro and modified or deleted the polygons that were inconsistent or inaccurate (249 out of 921 polygons. See Fig. 3 for examples), and manually digitized additional training areas following the same consistent LCZ physical definitions. For the two metropolitan areas where the expert dataset had no observations (Boston and San Francisco), training areas were digitized from scratch and used for both sets. Mturk-sourced polygons in the crowdsourced TA set were not used in the refined TA set.

2.2.2. Earth observation data

Landsat surface reflectance products were used as input in the LCZ modeling process. To minimize the potential impacts from land surface transition and the inter-annual variation of vegetation, seasonal composites for spring (April 1 to June 30), summer (July 1 to September 15),

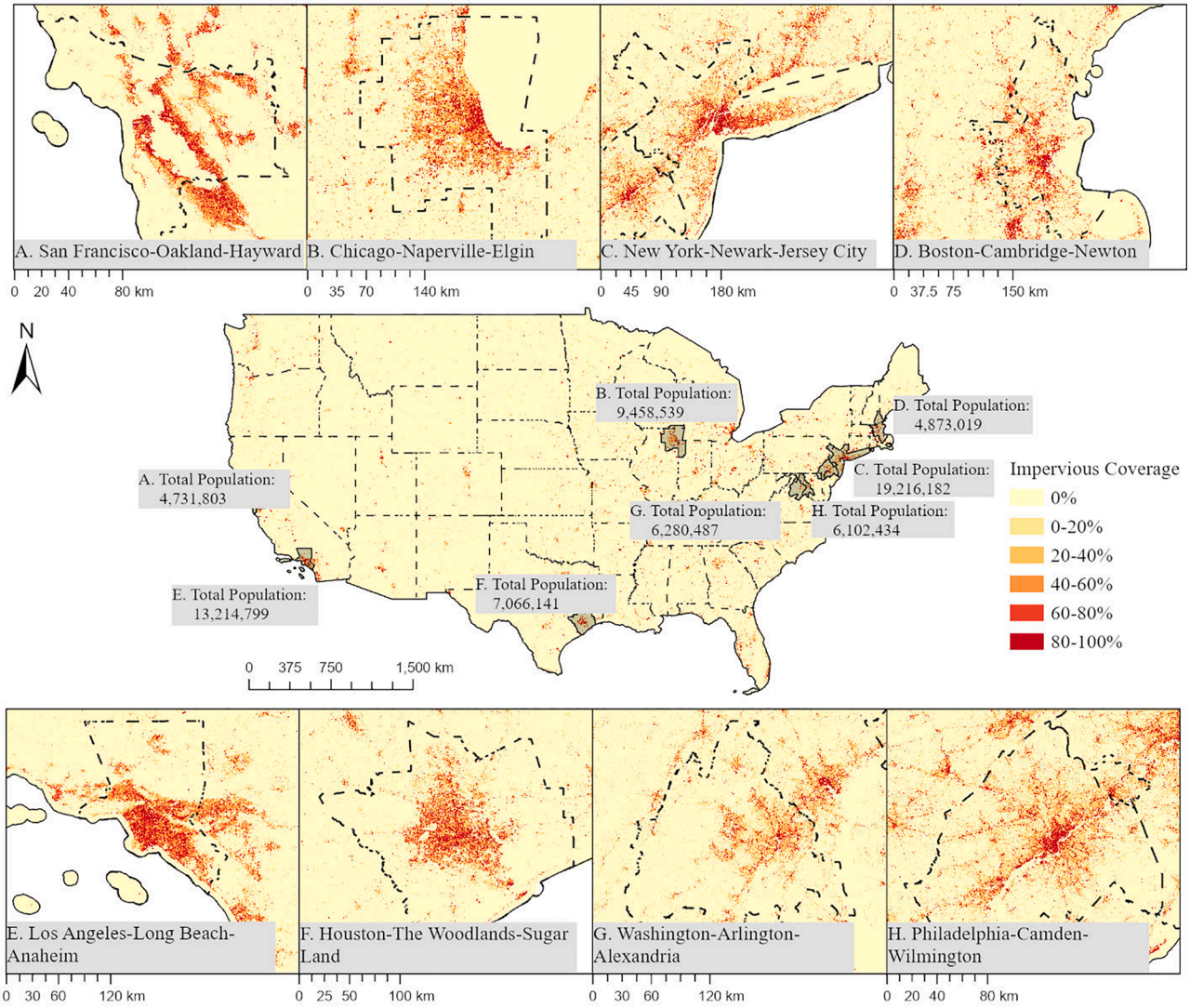


Fig. 2. Locations of the eight metropolitan areas. The urban layout is mapped with impervious surface coverage derived from NLCD 2016 Percent Developed Imperviousness (CONUS) (https://www.mrlc.gov/data).

and fall (September 15 to November 15) of the year 2018 were derived for each band and spectral index using the LandTrendr algorithm (see section 2.3.1).

2.3. Methods

2.3.1. Image processing

We used temporarily-stabilized imagery from the LandTrendr algorithm (Kennedy et al., 2010) as described in (Kennedy et al., 2018b) to generate predictors for our model. Temporal stabilization largely removes inter-annual signal noise in time-series imagery, allowing us to (1) interpolate observations in the year 2018 that are masked because of cloud and shadow, and (2) to remove some signal noise for the seasonal composites. We used 30 m spatial resolution Landsat surface reflectance products from Landsat 5 TM, Landsat 7 ETM+, and Landsat 8 OLI to build spectral data cubes from 1990 to 2018 for spring, summer, and fall, and then used LandTrendr to build temporally-smoothed cubes of imagery for a variety of original Landsat bands and derived indices (Table 1). We used Google Earth Engine for the LandTrendr analysis (Kennedy et al., 2018a); more details are available from https://emapr.github.io/LT-GEE/.

2.3.2. Image classification

We used a Random Forest (RF) classifier (Breiman, 2001) and a residual convolutional network (ResNet, He et al., 2016) for LCZ classification. RF is an ensemble classifier that produces multiple independent decision trees through bootstrapping with a randomly selected subset of training data and variables (Breiman, 2001). In this study, the six spectral bands and eight derived indices (Table 1) for each season were resampled to 100 m to get a 42-dimensional feature space, following the default LCZ mapping resolution suggested by the WUDAPT workflow. Then pixels with 42 spectral features were sampled and labeled within training areas and used as input for model training. The training process was implemented using the Scikit-learn Python module (https://scikit-learn.org/stable/) with default parameters provided by the module.

ResNet is a deep neural network that has shown superior performance in image classification tasks over traditional supervised classification algorithms. A ResNet reformulates the layers as learning residual functions with reference to the layer inputs (He et al., 2016), and it has shown good performance in LCZ mapping (Qiu et al., 2018, 2019). In this study, the implementation of a simple ResNet follows the practice of Qiu et al. (2018). Satellite imagery was resampled to 10 m, allowing 10*10 pixels to be placed in a 100 m LCZ grid (Yoo et al., 2019). We choose the patch size of 16*16 (corresponding to a minimum area of 160

Fig. 3. Examples to show the uncertainty in crowdsourced TAs. Fig. A and B are areas with similar surface properties (e.g., abundant vegetation and open arrangement of low-rise buildings), but were labeled as LCZ3 and LCZ6, respectively. Fig. C, D and E show training areas with highly heterogeneous LCZs within delineated polygons.

Table 1
LandTrendr (LT) parameters used in this study (Kennedy et al., 2018a).

LT Parameter	Type	LT Control Set
Spectral band		B; G; R; NIR; SWIR 1; SWIR 2
Spectral index		BU (He et al., 2010); NBR (Key and Benson, 2006); NDMI (Gao, 1996); NDVI (Rouse et al., 1974); NDWI (McFEETERS, 1996); TCB, TCG, TCW (Crist and Cicone, 1984)
maxSegments	Integer	8
spikeThreshold	Float	0.9
vertexCountOvershoot	Integer	3
preventOneYearRecovery	Boolean	True
RecoveryThreshold	Float	0.25
pvalThreshold	Float	0.05
bestModelProportion	Float	0.75
minObservationsNeeded	Integer	8

* 160 m^2, Fig. 4) for the input size of the ResNet classifier, with a core of 10*10 sampled within the training areas, and the pixels of surrounding areas sampled to include neighborhood information (Yoo et al., 2019). In the prediction process, neighborhood pixels were masked and the 10 * 10 core areas were labeled and resampled to 100 m following the default LCZ mapping resolution. The training process was implemented using the TensorFlow v2.0 framework (https://www.tensorflow.org/).

2.3.3. *The modeling process with four training area application schemes*

We tested two approaches to derive our samples. In the first approach, samples were drawn from all training areas, and subsequently 70% of samples were randomly selected for training and 30% for testing. We refer to this as "splitting the sample pool." In the second approach, the polygons themselves were first randomly split into 70/30 training/testing groups, and samples were drawn within them. We refer to this as "splitting the polygon pool." Splitting the polygon pool will lead to fewer training and testing samples being drawn nearby each other. For both approaches, we limited the number of training areas in some over-represented LCZ classes (such as LCZ6) to allow for a more balanced

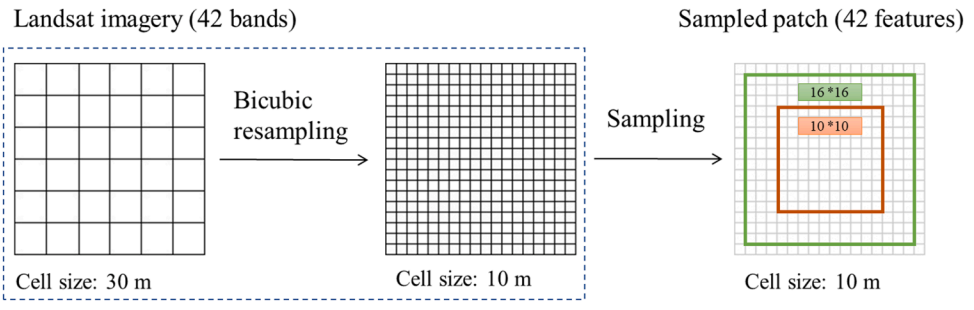


Fig. 4. The sampling workflow to get 16*16*42 input data for the ResNet.

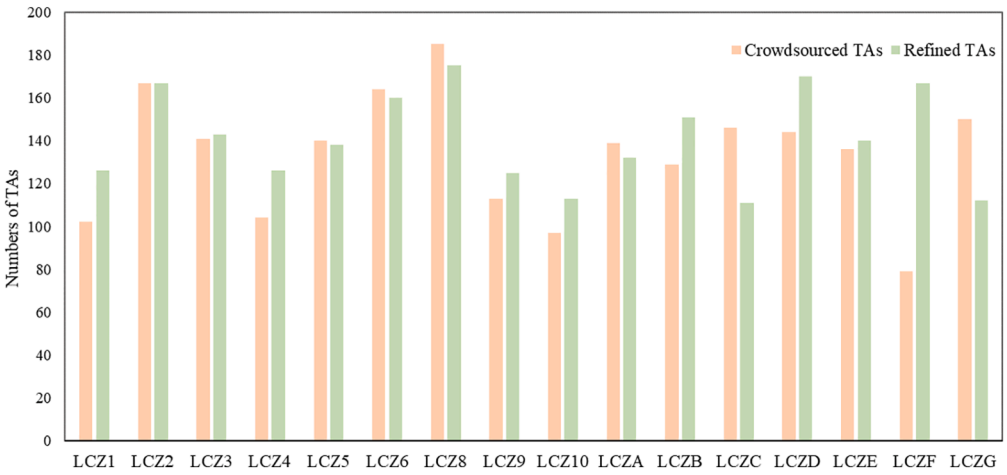


Fig. 5. The number of crowdsourced and refined TAs for each LCZ class.

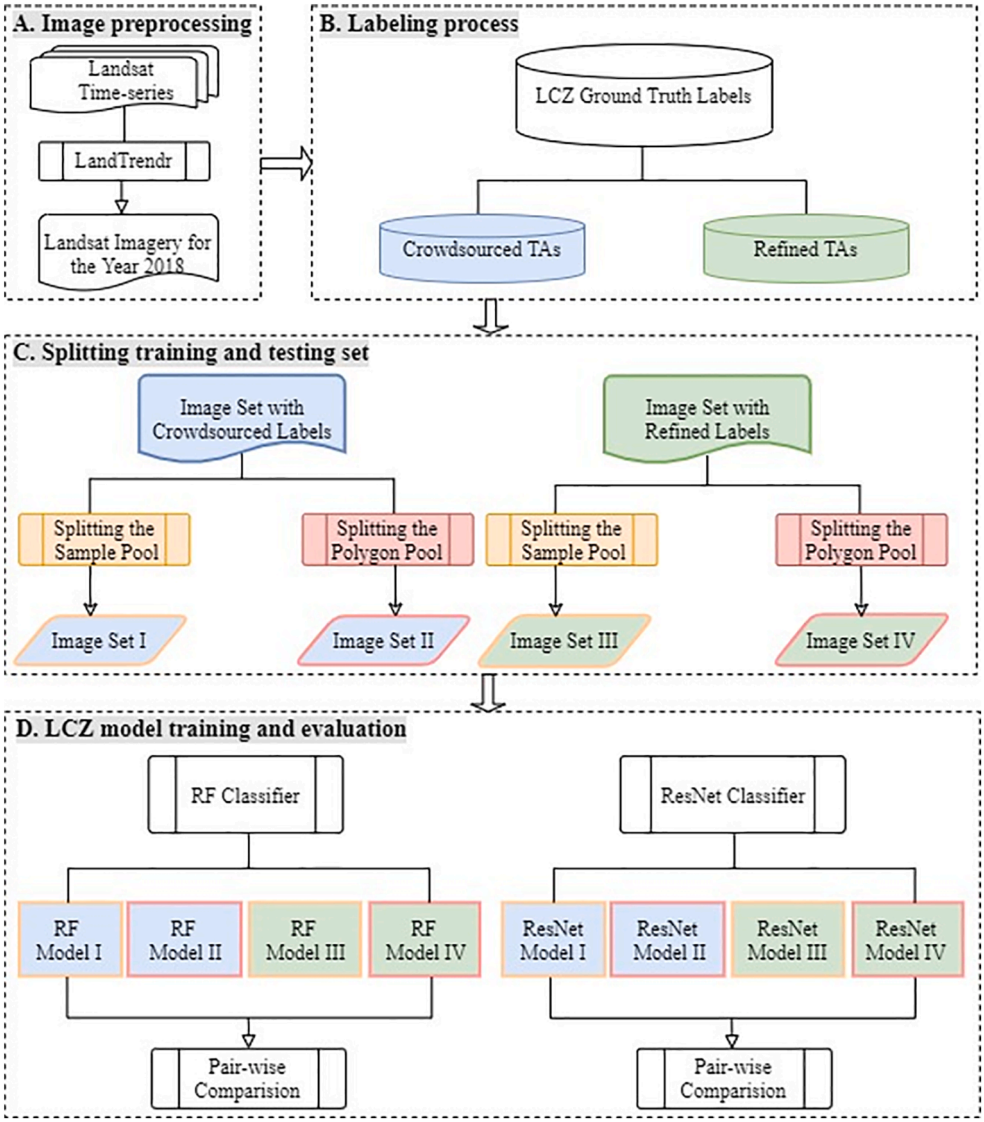


Fig. 6. The workflow of this study. Two training datasets were sampled from crowdsourced (blue color) and refined TAs (green color) at Stages A-B. Each of both was split by the sample pool (orange color) or polygon pool (red color) for training and testing at Stage C. Then four derived datasets were fed into RF and ResNet classifiers for LCZ classification and evaluation.

training and testing dataset across LCZs (Fig. 5).

For each classifier, there are four schemes for training area application: (1) sampling within the crowdsourced TAs and splitting the sample pool for training and testing purposes; (2) sampling within the crowdsourced TAs and splitting the polygon pool; (3) sampling within the refined TAs and splitting the sample pool; (4) sampling within the refined TAs and splitting the polygon pool. The model performance on the four schemes was compared for both classifiers. For each scheme, the training and testing splits were repeated 10 times for model evaluation, and model performances were compared across different schemes. The following metrics were computed for accuracy assessment: (1) Overall accuracy (OA): the percentage of correctly classified samples; (2) OA_u: the overall accuracy among the urban LCZ types (LCZ1–LCZ10); (3) OA_{bu}: the accuracy for the urban type after reclassifying the LCZ map to a binary map (natural and urban); (4) Weighted accuracy (WA): the accuracy after applying a similarity matrix to give different weights to different misclassifications (Bechtel et al., 2017). Additionally, F1-scores were obtained from the confusion matrix to further examine the class-wise accuracy. The F1-score is the harmonic mean of the user's accuracy (UA) and the producer's accuracy (PA) (Sokolova and Lapalme, 2009) and can take data imbalance into account. The workflow of this study is shown in Fig. 6.

$$F1 = (2*UA*PA)/(UA + PA) \quad (1)$$

3. Results

The two classifiers showed similar performances across four training area application schemes (Fig. 7). Overall, higher model accuracies were obtained on the refined TAs than on the crowdsourced TAs, and spatial autocorrelation boosted the apparent accuracy when the training and

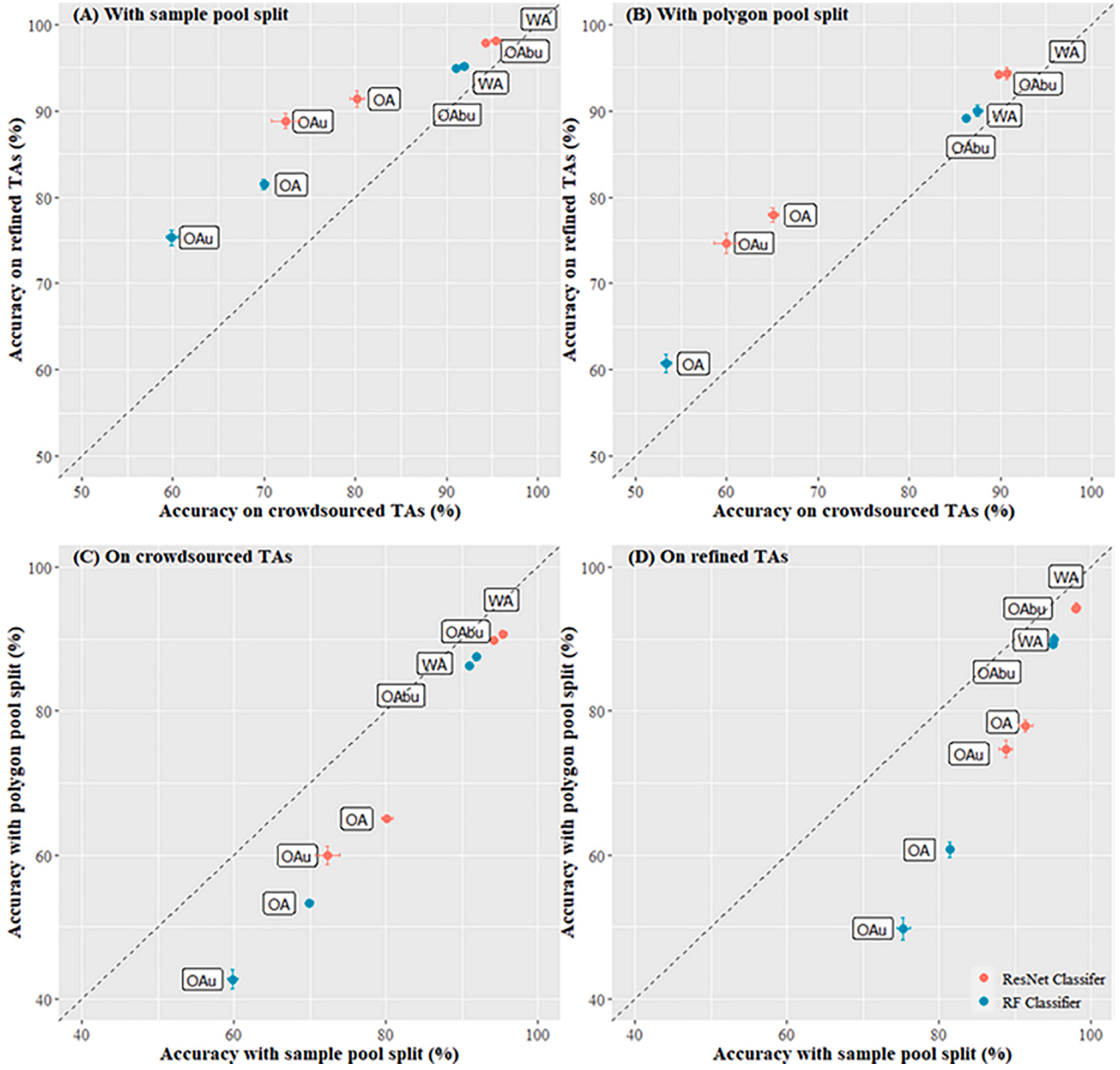


Fig. 7. Pair-wise comparison of the accuracy metrics for the four schemes on the RF and ResNet classifiers. The average values of accuracy metrics (OA: overall accuracy; OA_u: OA among the urban types; OA_{bu}: OA for the urban type on the reclassified binary map; WA: weighted accuracy) with standard deviations over 10 runs are shown with points and error bars. The points above or below the 1:1 line show the model accuracy of one scheme (on the y-axis) is higher or lower than another scheme (on the x-axis). For example, Fig. (A) shows that higher accuracies were obtained on the refined TAs compared to on the crowdsourced TAs when splitting the sample pool.

testing samples shared the same set of polygons.

3.1. RF classification results

The performance of the RF classifier varied on the four schemes. Averaged increases of 11.4% and 8.0% in overall accuracy were observed when using refined TAs compared to crowdsourced TAs on sample-pool split and polygon-pool split strategies, respectively. This indicates that discrepancies of LCZ labeling by different experts impact LCZ mapping, and the model performance can be potentially improved by refining the training areas. Average decreases of 17.2% and 20.8% were shown when splitting the polygons for the crowdsourced TAs and refined TAs, compared to the modeling process with splitting the sample pools. This further demonstrates the extent to which accuracy metrics could be inflated if the samples for training and testing originate in the same polygons.

The impact of training area application can be further shown in the class-wise classification accuracy in F1-scores and confusion matrices (Figs. 8-9). In general, the class-wise classification shows lower accuracy and larger variability for urban classes (LCZ1 to LCZ10) compared to natural classes (LCZA to LCZG). The F1-scores for the urban classes are low on the crowdsourced TAs, varying from 18% (LCZ4, Open high-rise) to 58% (LCZ6, Open low-rise) with polygon-pool split. The F1-scores for LCZ types were generally higher on the refined TAs than on the crowdsourced TAs. On polygon-pool split, the F1-score was improved by 9.5% on average for the urban classes using the refined TAs compared to the crowdsourced TAs. Significant improvements could be observed for LCZ1 (Compact high-rise, 13.6%), LCZ4 (Open high-rise, 16.0%), LCZ5 (Open midrise, 12.3%), and LCZ9 (Sparsely built, 18.3%). The F1-score was improved by 13.6% on average for natural types, with significant improvements shown for LCZB (Scattered trees, 17.5%), LCZC (Bush/scrub, 15.1%), LCZD (Low plants, 20.2%), and LCZF (Bare soil or sand, 26.0%). These LCZ types were easily misclassified (Fig. 9), but the misclassification could be mitigated by refining training areas (see Fig. 10).

Additionally, F1-scores were inflated when training and testing pixels shared the sample polygons, especially for the urban types. On the refined TAs, the F1-score decreased by 26.9% on average for the urban

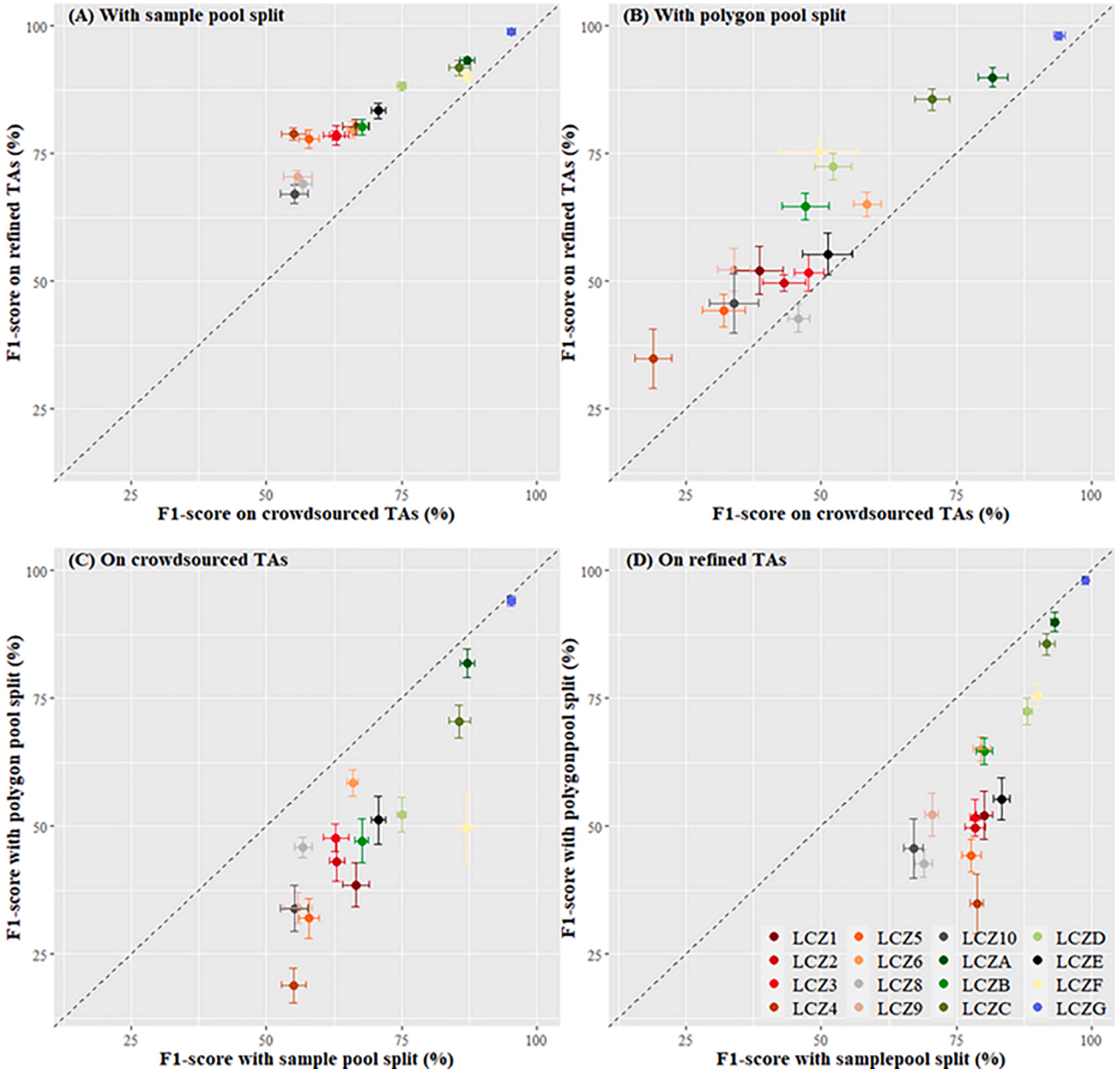


Fig. 8. The pairwise comparison of F1-scores for the four schemes on the RF classifier.

(a) sampling within the crowdsourced TAs and splitting the sample pool

	LCZ1	LCZ2	LCZ3	LCZ4	LCZ5	LCZ6	LCZ8	LCZ9	LCZ10	LCZA	LCZB	LCZC	LCZD	LCZE	LCZF	LCZG	PA%
LCZ1	189	45	11	9	13	6	13	0	4	0	0	0	0	12	0	0	63%
LCZ2	28	333	45	6	20	9	21	0	10	0	0	0	0	9	1	0	69%
LCZ3	2	27	312	8	11	47	13	1	7	0	1	0	1	3	0	0	72%
LCZ4	27	44	24	137	24	20	22	2	2	0	6	0	1	3	0	2	44%
LCZ5	8	39	40	20	265	28	28	6	3	1	4	0	0	9	0	0	59%
LCZ6	0	4	31	3	15	346	19	25	0	27	13	4	2	0	0	1	71%
LCZ8	8	38	40	5	51	32	314	12	22	3	5	0	11	24	3	0	55%
LCZ9	1	0	1	0	5	29	10	175	2	31	29	6	32	7	13	2	51%
LCZ10	11	22	18	2	8	5	46	2	125	0	1	0	1	26	2	3	46%
LCZA	0	0	0	0	0	1	0	3	0	383	9	1	2	0	0	0	96%
LCZB	0	1	4	2	4	13	0	19	0	23	250	25	24	1	11	0	66%
LCZC	0	1	1	0	0	7	4	8	0	5	11	406	9	10	6	0	87%
LCZD	0	1	0	1	1	8	3	11	0	7	31	10	314	5	15	3	77%
LCZE	11	14	7	0	5	8	25	1	20	1	0	10	1	279	5	0	72%
LCZF	0	0	1	0	0	0	5	0	3	0	0	2	10	10	355	1	92%
LCZG	0	1	1	1	1	1	1	3	1	2	3	0	5	3	1	422	95%
UA%	66%	58%	58%	71%	63%	62%	60%	65%	63%	79%	69%	88%	76%	70%	86%	97%	

(b) sampling within the crowdsourced TAs and splitting the polygon pool

	LCZ1	LCZ2	LCZ3	LCZ4	LCZ5	LCZ6	LCZ8	LCZ9	LCZ10	LCZA	LCZB	LCZC	LCZD	LCZE	LCZF	LCZG	PA%
LCZ1	76	30	10	9	4	13	15	1	7	0	0	0	0	10	0	4	42%
LCZ2	37	185	54	7	24	7	38	0	6	0	0	0	0	3	0	0	51%
LCZ3	1	59	192	0	16	70	32	5	12	0	2	0	1	20	0	0	47%
LCZ4	32	43	25	32	32	31	41	4	4	0	5	0	1	6	0	5	12%
LCZ5	7	34	48	25	98	62	69	7	6	0	3	1	3	6	2	1	26%
LCZ6	0	0	25	3	19	342	19	16	1	35	30	2	3	4	1	0	68%
LCZ8	12	62	30	6	29	43	276	7	24	3	6	1	5	28	13	3	50%
LCZ9	1	1	4	0	6	70	14	96	1	30	45	10	39	3	12	2	29%
LCZ10	11	27	42	0	6	5	83	1	60	0	2	0	3	41	5	5	21%
LCZA	0	0	0	0	0	5	0	3	0	406	5	1	0	0	0	0	97%
LCZB	0	1	6	5	1	23	2	29	0	31	197	39	24	3	19	1	52%
LCZC	0	0	0	0	1	17	5	19	0	21	41	284	12	22	8	1	66%
LCZD	0	0	0	0	2	21	7	43	0	6	100	9	200	3	31	1	47%
LCZE	26	11	9	4	5	8	35	0	27	0	0	7	9	198	24	0	55%
LCZF	2	0	0	0	0	0	8	0	9	0	3	6	31	15	127	2	63%
LCZG	1	3	0	4	1	1	1	4	1	4	1	0	4	4	12	408	91%
UA%	37%	41%	43%	34%	40%	48%	43%	41%	38%	76%	45%	79%	60%	54%	50%	94%	

(c) sampling within the refined TAs and splitting the sample pool

	LCZ1	LCZ2	LCZ3	LCZ4	LCZ5	LCZ6	LCZ8	LCZ9	LCZ10	LCZA	LCZB	LCZC	LCZD	LCZE	LCZF	LCZG	PA%
LCZ1	275	20	22	12	7	0	9	0	10	0	0	0	0	6	2	0	76%
LCZ2	21	418	21	7	10	0	14	0	4	0	0	0	0	5	2	0	83%
LCZ3	0	23	357	8	5	9	7	1	2	0	0	0	0	0	0	0	87%
LCZ4	5	13	16	323	37	10	5	0	0	0	2	0	0	1	0	0	78%
LCZ5	1	12	14	23	424	36	5	3	0	0	0	0	1	0	0	0	82%
LCZ6	0	6	17	1	28	465	3	24	0	2	2	0	1	0	2	0	84%
LCZ8	8	30	27	4	21	7	331	2	29	0	0	0	1	25	7	0	67%
LCZ9	0	0	0	0	3	41	3	253	0	23	22	8	12	0	5	0	68%
LCZ10	8	22	7	0	6	0	44	2	193	0	0	0	0	36	10	2	58%
LCZA	0	0	0	0	0	0	0	7	0	397	6	3	0	0	0	0	96%
LCZB	0	0	0	1	0	6	0	27	0	8	342	21	29	0	6	0	78%
LCZC	0	0	0	0	0	0	0	2	0	0	10	320	0	0	4	0	95%
LCZD	0	0	0	0	0	3	0	7	0	2	18	4	418	0	15	0	90%
LCZE	5	5	12	1	2	0	14	0	14	0	0	0	0	356	3	0	86%
LCZF	1	0	0	0	1	0	8	3	5	0	6	4	6	3	398	0	91%
LCZG	0	0	0	1	1	0	0	1	0	4	0	0	0	0	0	341	98%
UA%	85%	76%	72%	85%	78%	81%	75%	76%	75%	91%	84%	89%	89%	82%	88%	99%	

(d) sampling within the refined TAs and splitting the polygon pool

	LCZ1	LCZ2	LCZ3	LCZ4	LCZ5	LCZ6	LCZ8	LCZ9	LCZ10	LCZA	LCZB	LCZC	LCZD	LCZE	LCZF	LCZG	PA%
LCZ1	190	76	22	16	7	1	24	0	17	0	0	0	0	17	3	0	51%
LCZ2	54	258	66	6	26	2	61	0	12	0	0	0	0	20	5	0	51%
LCZ3	5	48	230	10	37	40	24	3	15	0	0	0	0	13	5	0	53%
LCZ4	9	31	36	114	85	53	25	12	6	0	6	0	0	2	0	1	30%
LCZ5	7	22	43	51	230	106	36	23	1	0	3	0	4	0	4	0	43%
LCZ6	0	5	37	3	34	376	4	52	2	2	5	0	0	0	0	0	72%
LCZ8	6	42	51	6	34	18	236	6	52	0	0	0	0	37	12	1	47%
LCZ9	0	0	0	1	4	61	0	219	4	35	31	10	8	0	7	0	58%
LCZ10	10	20	24	3	5	2	65	2	135	0	1	0	0	63	10	0	40%
LCZA	0	0	0	0	0	0	0	6	0	385	6	3	0	0	0	0	96%
LCZB	0	0	0	1	1	7	0	35	0	15	324	31	33	0	13	0	70%
LCZC	0	0	0	0	0	1	0	3	0	5	12	300	5	0	14	0	88%
LCZD	0	0	0	0	5	1	4	19	0	3	112	3	311	0	12	0	66%
LCZE	12	15	11	4	6	0	62	0	57	0	0	0	0	231	20	2	55%
LCZF	0	6	15	0	3	8	25	9	19	0	1	8	20	3	353	0	75%
LCZG	0	1	0	0	0	0	4	1	0	0	0	0	0	0	0	333	98%
UA%	65%	49%	43%	53%	48%	56%	41%	56%	42%	87%	65%	85%	82%	60%	77%	99%	

Accuracy (%)
0 20 40 60 80 100

Fig. 9. Confusion matrices of the model with highest overall accuracy among 10 runs on the RF.

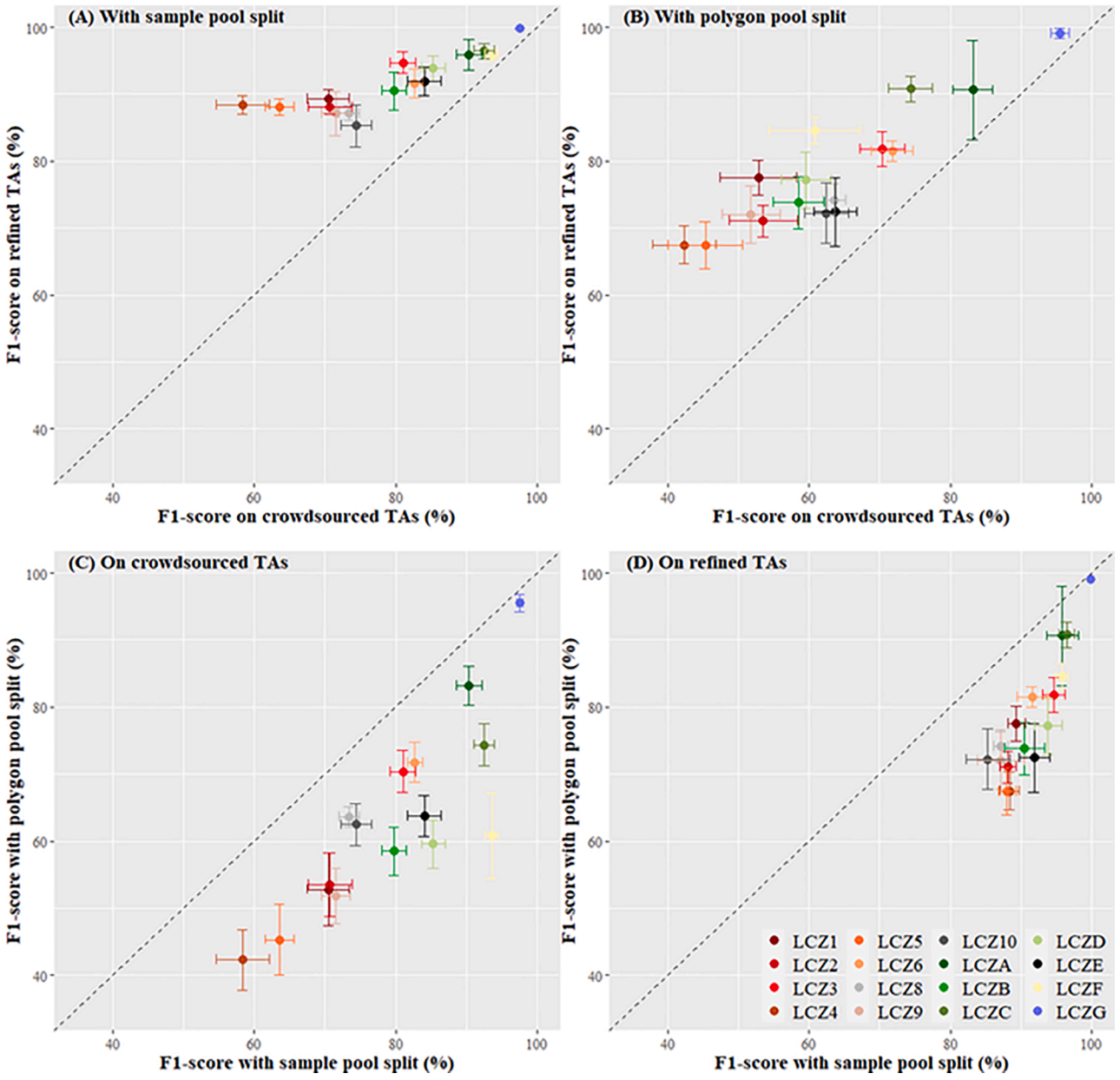


Fig. 10. The pairwise comparison of F1-scores for the four schemes on the ResNet.

classes with polygon-pool split compared to the sample-pool split. The highest drop in F1-score was observed for LCZ4 (43.9%). Large drops in F1-scores (>20%) also showed for compact and some open/low-rise types (LCZ1, LCZ2, LCZ3, LCZ5, and LCZ8). Additionally, F1-score was decreased by 12.1% on average for the natural classes. Similar results were shown on the crowdsourced TAs. With polygon-pool split, the averaged F1-scores decreased from 60.0% to 39.1% for urban types, and from 81.2% to 63.7% for natural types.

3.2. ResNet classification results

Similar classification results were shown on the ResNet. On the refined TAs, the overall accuracies increased by 11.5% and 7.4% on the sample-split and polygon-split strategies, respectively, compared to the results on the crowdsourced TAs. Average decreases of 16.6% and 20.7% were shown on polygon-pool split for the crowdsourced and refined TAs respectively, compared to sample-pool split. With polygon held-out, the F1-scores on refined TAs were improved by 15.9% and 13.3% on average for the urban class and natural class respectively compared to the crowdsourced TAs. Significant classification improvements (>20%) were obtained for several urban classes, including LCZ1, LCZ4, LCZ5, and LCZ9. Similarly, using the refined TAs, the F1-scores decreased by 14.9% and 10.8% on average for the urban and natural classes, respectively, when sampling from the split polygons instead of the sample pool. The large drops in F1-scores were observed for LCZ4 (20.9%) and LCZ5 (20.6%). Fig. 11 shows the ResNet mapping results using the refined TAs and the polygon pool split strategy.

3.3. Discrepancies of LCZ maps beyond accuracy metrics

The confusion matrices show that spatial autocorrelation can boost accuracy metrics on different training sets and classifiers, which can be further shown in the resulting LCZ maps. Consider the Chicago urban area as an example (Fig. 12). The output map from the ResNet is less patchy, indicating that the zones were better predicted in the complete shape and boundary conditions. On the refined TAs, the ResNet classifier

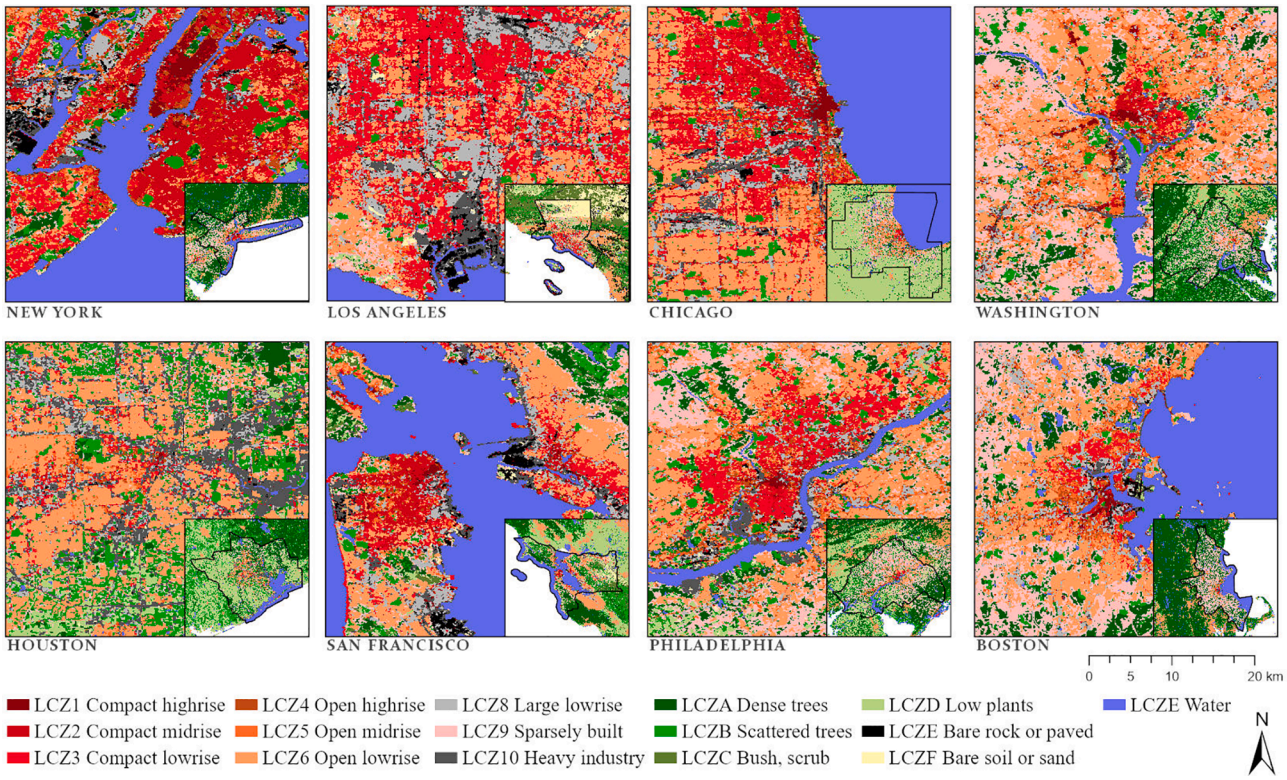


Fig. 11. LCZ maps from the ResNet. The refined TAs were split on the polygon pool.

achieved 79.5% in predicted accuracy for LCZ1 on polygon-pool split, similar to the prediction accuracy on the RF classifier with sample-pool split (80.1%), whereas the LCZ maps showed large discrepancies (See Fig. 12 for the example in the zoomed-in downtown areas). Furthermore, with sample-pool split, the RF classifier had a lower predicted accuracy (64.4%) for LCZ1 on crowdsourced TAs compared to refined TAs, but the LCZ maps showed similar spatial patterns for the two schemes. Thus, spatial autocorrelation in the training and testing sets boosts the apparent classification accuracy but fails to result in better LCZ maps. Similar examples can also be found for other zones (Fig. 12).

4. Discussion

Training areas collected for worldwide cities have been widely used in LCZ classification, especially using the Landsat archive at medium resolution. As digitized polygons are the source of input data to machine learning classifiers, training area collection and usage is critical. Moreover, broad-scale LCZ mapping is more challenging compared to mapping an individual city.

The identification of training areas is a difficult and time-demanding task. The LCZ scheme was not initially developed for mapping based on the spectral properties of geographic zones on satellite imagery. Instead, the 17 classes were differentiated according to many factors including land cover composition, urban fabric, and functional use. The geometric and surface cover properties distinguishing LCZ types (Fig. 1) are often difficult to infer from high-resolution imagery alone. For example, LCZ4 is defined with 20–40% building coverage, 30–40% impervious surface coverage, and 30–40% pervious surface coverage, but those properties could not be extracted in a straightforward manner during the human interpretation from high-resolution images. Some urban types are difficult to distinguish without the help of ancillary data sources. For instance, LCZ4 and LCZ5 have similar surface fractions but differ in the height of roughness elements (>25 m for LCZ4 and 3–10 m for LCZ5). These two open types are difficult to label if 3D urban structural data such as Google Street View imagery are unavailable. Furthermore, although the LCZ scheme has been generally accepted as a culturally-neutral classification scheme for global cities, the diverse pattern of the urban layout means that internal heterogeneity is unavoidably inherent in each class. The urban landscape is a continuum of spatial pattern and surface properties. The existence of unclear or fuzzy boundaries makes it difficult to divide a whole city into the LCZ classes with high certainty, especially at the boundary of two discrete LCZ types. The bias in human interpretation of the LCZ scheme is far more pronounced when the training areas for different cities are sampled by different experts.

Crowdsourced TAs vary in level of quality and should be used with more caution in broad-scale LCZ mapping. The inconsistency of the crowdsourced TAs induced by bias in human interpretation can lead to erroneous mapping results and poor model performance, as is highlighted in the disagreement with implementing the crowdsourced and refined TAs in this study. Furthermore, the predictive power of the classifier can be overestimated in broad-scale LCZ mapping when spatial autocorrelation structures in the performance estimation setting do not match those of the ultimate predictive setting. Training and testing sets should be independent in the case that the model will be applied to input data unseen in the training set from the real world.

Our examination of these challenges gives insight into improving community science initiatives by addressing concerns about data quality and usage. Given that contributors vary in expertise and may collect data that do not satisfy the protocols, innovative workflows for the labeling process to digitize training areas or the raw training dataset should be explored. In the current workflow, having one or a few well-trained experts to check or vote the crowdsourced data as a quality control can potentially improve the data quality. With the development of sophisticated deep learning algorithms, automatic detection and segmentation of Google Street View images and high-resolution imagery can provide urban parameters such as building heights that are difficult for human experts to perceive from images. Additionally, cloud-computing platforms such as Google Earth Engine largely reduce the burdensome pre-processing requirements of satellite image analysis,

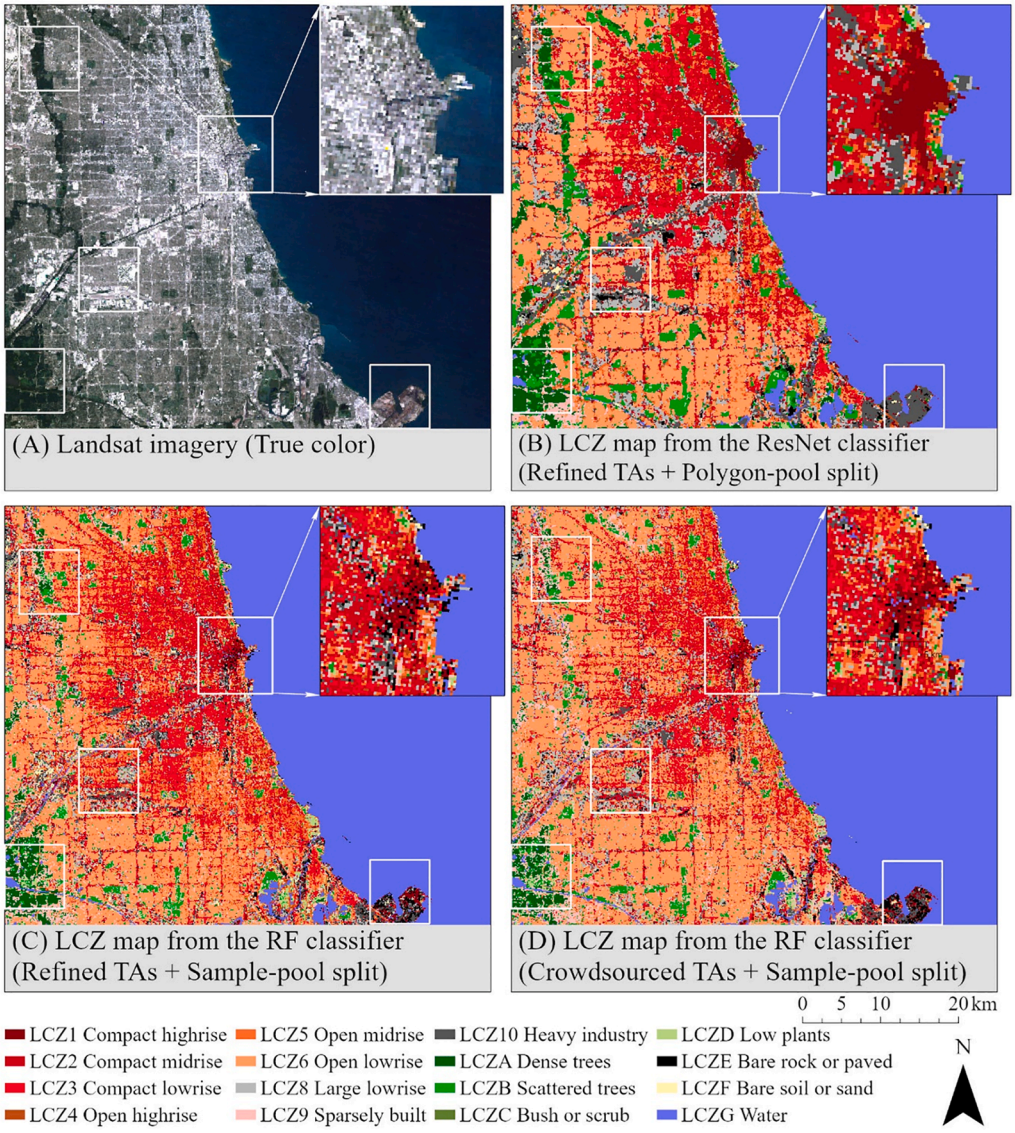


Fig. 12. LCZ maps for the Chicago area.

making it possible to provide more targeted instructions on data collection to ensure quality. With powerful and efficient computing resources, the contribution and involvement of local communities from worldwide cities can be realized in a more collaborative way for broad-scale LCZ mapping.

Our study supports the finding of other researchers that model accuracy can be overestimated if the spatial dependency in training data is ignored in the modeling process. Ploton et al. (2020) argued that highly optimistic evaluation is a common problem in big-data mapping practices, and illustrated their argument with biomass mapping in Africa. Roberts et al. (2017) showed that ignoring spatial structure in ecological data leads to underestimation of model predictive error when performing cross-validation. In recent years, some studies have demonstrated efforts to reduce the impact of spatial autocorrelation in the data sampling process for LCZ mapping. For instance, Zhu et al. (2020) developed a valuable benchmark dataset, So2Sat LCZ42, for global LCZ classification, which consists of a large number of image patches (dimensions 320 m * 320 m) collected and labeled from Sentinel imagery. The validation and testing sets were separated by a half-city held-out approach, and the training separated from both sets geographically with city held-out splitting. This dataset has been applied in LCZ model development and practical mapping (Qiu et al., 2020). Another example comes from Rosentreter et al. (2020), where the WUDAPT workflow was followed but training and testing areas were separated by a minimum distance of 320 m. Our findings suggest that further work is justified to more explicitly examine the spatial structure of training data selection. Two lines of further study appear reasonable. First, characterization of within-polygon spatial autocorrelation, perhaps split by different LCZ types, could help determine if the range of spatial variation within large training areas could support within-polygon sampling. Second, characterization of cross-polygon autocorrelation at different scales could be useful to determine if further autocorrelation remains even with polygon-splitting.

Finally, our results suggest that the LandTrendr temporal-stabilization process provided imagery that resulted in classification model comparable to other studies (Bechtel et al., 2019; Ren et al., 2019). Thus, temporally-smoothed Landsat imagery has the potential to be used for multi-year LCZ mapping in future applications, which can help characterize long-term urban dynamics in the land surface and local climate.

5. Conclusions

In this study, we explored two issues associated with training areas

serving as ground-truth labels for broad-scale LCZ mapping. One is the impact of the inconsistency in crowdsourced labeling process by different interpreters, and the other is overestimation of the model performance due to spatial autocorrelation in the training data. By applying two different classifiers to map eight metropolitan areas in the US into LCZs, we found that both RF and ResNet show poorer prediction performances on crowdsourced TAs compared to on refined TAs, especially for urban types. Overall accuracies were 54% and 64% on the RF and ResNet using the crowdsourced TAs, respectively, after the impact of spatial autocorrelation was reduced by polygon held-out approach. However, spatial autocorrelation can boost the apparent accuracy of the classifiers by 16% to 21%. Optimistic accuracy can obscure the inadequate quality of crowdsourced TAs and lead to erroneous interpretation of mapping results. With the interaction of the two issues, we argue that the uncertainty in the crowdsourced labeling process and data annotation accuracy deserves more attention in broad-scale LCZ mapping studies. Additionally, mapping workflows need to consider differences in spatial autocorrelation structures among training, validation, and testing sets.

Funding: This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

CRediT authorship contribution statement

Chunxue Xu: Conceptualization, Methodology, Data curation, Formal analysis, Writing – original draft, Writing – review & editing. **Perry Hystad:** Conceptualization, Methodology, Supervision, Writing – review & editing. **Rui Chen:** Formal analysis, Validation. **Jamon Van Den Hoek:** Writing – review & editing. **Rebecca A. Hutchinson:** Writing – review & editing. **Steve Hankey:** Data curation, Writing – review & editing. **Robert Kennedy:** Conceptualization, Methodology, Supervision, Writing – review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Agathangelidis, I., Cartalis, C., Santamouris, M., 2019. Integrating Urban Form, Function, and Energy Fluxes in a Heat Exposure Indicator in View of Intra-Urban Heat Island Assessment and Climate Change Adaptation. Climate 7 (6), 75. https://doi.org/10.3390/cli7060075.

Bechtel, B., Alexander, P.J., Beck, C., Böhner, J., Brousse, O., Ching, J., Demuzere, M., Fonte, C., Gál, T., Hidalgo, J., Hoffmann, P., Middel, A., Mills, G., Ren, C., See, L., Sismanidis, P., Verdonck, M.-L., Xu, G., Xu, Y., 2019. Generating WUDAPT Level 0 data – Current status of production and evaluation. Urban Climate 27, 24–45.

Bechtel, B., Alexander, P., Böhner, J., Ching, J., Conrad, O., Feddema, J., Mills, G., See, L., Stewart, I., 2015. Mapping Local Climate Zones for a Worldwide Database of the Form and Function of Cities. IJGI 4 (1), 199–219.

Bechtel, B., Demuzere, M., Sismanidis, P., et al, 2017. Quality of Crowdsourced Data on Urban Morphology—The Human Influence Experiment (HUMINEX) 21.

Breiman, L., 2001. Random Forests. Machine Learning 45, 5–32.

Crist, E.P., Cicone, R.C., 1984. A Physically-Based Transformation of Thematic Mapper Data—The TM Tasseled Cap. IEEE Trans. Geosci. Remote Sensing GE-22, 256–263.

Demuzere, M., Bechtel, B., Middel, A., Mills, G., Mourshed, M., 2019. Mapping Europe into local climate zones. PLoS ONE 14 (4), e0214474. https://doi.org/10.1371/journal.pone.0214474.

Demuzere, M., Hankey, S., Mills, G., Zhang, W., Lu, T., Bechtel, B., 2020a. Combining expert and crowd-sourced training data to map urban form and functions for the continental US. Sci Data 7 (1). https://doi.org/10.1038/s41597-020-00605-z.

Demuzere, M., Hankey, S., Mills, G., Zhang, W., Lu, T., Bechtel, B., 2020b. CONUS-wide LCZ map and Training Areas. figshare. Dataset. https://doi.org/10.6084/m9.figshare.11416950.v2.

Dhar, T.K., Khirfan, L., 2017. A multi-scale and multi-dimensional framework for enhancing the resilience of urban form to climate change. Urban Climate 19, 72–91.

Esch, T., Heldens, W., Hirner, A., Keil, M., Marconcini, M., Roth, A., Zeidler, J., Dech, S., Strano, E., 2017. Breaking new ground in mapping human settlements from space – The Global Urban Footprint. ISPRS Journal of Photogrammetry and Remote Sensing 134, 30–42.

Forman, R.T.T., 2014. Urban Ecology: Science of Cities. Cambridge University Press.

Gao, B.-C., 1996. NDWI—A normalized difference water index for remote sensing of vegetation liquid water from space. Remote Sensing of Environment 58 (3), 257–266.

He, C., Shi, P., Xie, D., Zhao, Y., 2010. Improving the normalized difference built-up index to map urban built-up areas using a semiautomatic segmentation approach. Remote Sensing Letters 1 (4), 213–221.

He, K., Zhang, X., Ren, S., et al., 2016. Deep Residual Learning for Image Recognition. in: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

Kennedy, R., Yang, Z., Gorelick, N., Braaten, J., Cavalcante, L., Cohen, W., Healey, S., 2018a. Implementation of the LandTrendr Algorithm on Google Earth Engine. Remote Sensing 10 (5), 691. https://doi.org/10.3390/rs10050691.

Kennedy, R.E., Ohmann, J., Gregory, M., Roberts, H., Yang, Z., Bell, D.M., Kane, V., Hughes, M.J., Cohen, W.B., Powell, S., Neeti, N., Larrue, T., Hooper, S., Kane, J., Miller, D.L., Perkins, J., Braaten, J., Seidl, R., 2018b. An empirical, integrated forest biomass monitoring system. Environ. Res. Lett. 13 (2), 025004. https://doi.org/10.1088/1748-9326/aa9d9e.

Kennedy, R.E., Yang, Z., Cohen, W.B., 2010. Detecting trends in forest disturbance and recovery using yearly Landsat time series: 1. LandTrendr — Temporal segmentation algorithms. Remote Sensing of Environment 114 (12), 2897–2910.

Key, C.H., Benson, N.C., 2006. Landscape Assessment: Ground measure of severity, the Composite Burn Index; and Remote sensing of severity, the Normalized Burn Ratio. USDA Forest Service, Rocky Mountain Research Station, Ogden, UT.

Marconcini, M., Metz-Marconcini, A., Üreyen, S., Palacios-Lopez, D., Hanke, W., Bachofer, F., Zeidler, J., Esch, T., Gorelick, N., Kakarla, A., Paganini, M., Strano, E., 2020. Outlining where humans live, the World Settlement Footprint 2015. Scientific Data 7 (1). https://doi.org/10.1038/s41597-020-00580-5.

McFEETERS, S.K., 1996. The use of the Normalized Difference Water Index (NDWI) in the delineation of open water features. International Journal of Remote Sensing 17 (7), 1425–1432.

Middel, A., Lukasczyk, J., Maciejewski, R., Demuzere, M., Roth, M., 2018. Sky View Factor footprints for urban climate modeling. Urban Climate 25, 120–134.

Ploton, P., Mortier, F., Réjou-Méchain, M., Barbier, N., Picard, N., Rossi, V., Dormann, C., Cornu, G., Viennois, G., Bayol, N., Lyapustin, A., Gourlet-Fleury, S., Pélissier, R., 2020. Spatial validation reveals poor predictive performance of large-scale ecological mapping models. Nature Communications 11 (1). https://doi.org/10.1038/s41467-020-18321-y.

Qiu, C., Mou, L., Schmitt, M., Zhu, X.X., 2019. Local climate zone-based urban land cover classification from multi-seasonal Sentinel-2 images with a recurrent residual network. ISPRS Journal of Photogrammetry and Remote Sensing 154, 151–162.

Qiu, C., Schmitt, M., Mou, L., Ghamisi, P., Zhu, X., 2018. Feature Importance Analysis for Local Climate Zone Classification Using a Residual Convolutional Neural Network with Multi-Source Datasets. Remote Sensing 10 (10), 1572. https://doi.org/10.3390/rs10101572.

Qiu, C., Tong, X., Schmitt, M., Bechtel, B., Zhu, X.X., 2020. Multilevel Feature Fusion-Based CNN for Local Climate Zone Classification From Sentinel-2 Images: Benchmark Results on the So2Sat LCZ42 Dataset. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 13, 2793–2806.

Ren, C., Cai, M., Li, X., Zhang, L., Wang, R., Xu, Y., Ng, E., 2019. Assessment of Local Climate Zone Classification Maps of Cities in China and Feasible Refinements. Scientific Reports 9 (1). https://doi.org/10.1038/s41598-019-55444-9.

Roberts, D.R., Bahn, V., Ciuti, S., Boyce, M.S., Elith, J., Guillera-Arroita, G., Hauenstein, S., Lahoz-Monfort, J.J., Schröder, B., Thuiller, W., Warton, D.I., Wintle, B.A., Hartig, F., Dormann, C.F., 2017. Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. Ecography 40 (8), 913–929.

Rosentreter, J., Hagensieker, R., Waske, B., 2020. Towards large-scale mapping of local climate zones using multitemporal Sentinel 2 data and convolutional neural networks. Remote Sensing of Environment 237, 111472. https://doi.org/10.1016/j.rse.2019.111472.

Rouse Jr., J.W., Haas, R.H., Schell, J.A., Deering, D.W., 1974. Monitoring Vegetation Systems in the Great Plains with Erts. NASA Special Publication 351, 309.

Seto, K.C., Sánchez-Rodríguez, R., Fragkias, M., 2010. The New Geography of Contemporary Urbanization and the Environment. Annu. Rev. Environ. Resour. 35 (1), 167–194.

Sokolova, M., Lapalme, G., 2009. A systematic analysis of performance measures for classification tasks. Information Processing & Management 45 (4), 427–437.

Stewart, I.D., Oke, T.R., 2012. Local Climate Zones for Urban Temperature Studies. Bull. Amer. Meteor. Soc. 93, 1879–1900.

United Nations, 2014. World urbanization prospects, the 2014 revision: highlights. United Nations. https://doi.org/10.18356/527e5125-en.

Xu, Y., Goodacre, R., 2018. On splitting training and validation set: A comparative study of cross-validation, bootstrap and systematic sampling for estimating the generalization performance of supervised learning. Journal of Analysis and Testing 2 (3), 249–262.

Yoo, C., Han, D., Im, J., Bechtel, B., 2019. Comparison between convolutional neural networks and random forest for local climate zone classification in mega urban areas using Landsat images. ISPRS Journal of Photogrammetry and Remote Sensing 157, 155–170.

Zhu, X.X., Hu, J., Qiu, C., Shi, Y., Kang, J., Mou, L., Bagheri, H., Haberle, M., Hua, Y., Huang, R., Hughes, L., Li, H., Sun, Y., Zhang, G., Han, S., Schmitt, M., Wang, Y., 2020. So2Sat LCZ42: A Benchmark Data Set for the Classification of Global Local Climate Zones [Software and Data Sets]. IEEE Geoscience and Remote Sensing Magazine 8 (3), 76–89.