Quality Assessment of Heterogeneous Training Data Sets for Classification of Urban Area with Landsat Imagery

Neema Nicodemus Lyimo, Fang Luo, Qimin Cheng, and Hao Peng

Abstract

Quality assessment of training samples collected from heterogeneous sources has received little attention in the existing literature. Inspired by Euclidean spectral distance metrics, this article derives three quality measures for modeling uncertainty in spectral information of open-source heterogeneous training samples for classification with Landsat imagery. We prepared eight test case data sets from volunteered geographic information and open government data sources to assess the proposed measures. The data sets have significant variations in quality, quantity, and data type. A correlation analysis verifies that the proposed measures can successfully rank the quality of heterogeneous training data sets prior to the image classification task. In this era of big data, preclassification quality assessment measures empower research scientists to select suitable data sets for classification tasks from available open data sources. Research findings prove the versatility of the Euclidean spectral distance function to develop quality metrics for assessing open-source training data sets with varying characteristics for urban area classification.

Introduction

Data quality is defined as a concept that includes data precision and accuracy to determine if the data are specific enough and the types or amount of errors they contain (Bielecka and Burek 2019). However, in this research, we are interested in a broader definition that embraces aspects of data relevance—data qualities that are often characterized as "fitness for use" (Stanislawski *et al.* 2014; Zhou *et al.* 2018; Shao *et al.* 2018). This article intends to assess the suitability of data extracted from multiple open-source platforms as training sets to classify or retrieve from remotely sensed images.

The increasing availability of crowdsourced data and other free open data sources brings new opportunities for geospatial applications (Deren *et al.* 2014, 2019; Yin *et al.* 2015; Shao *et al.* 2020). Despite its success, every source has its related challenges. For example, OpenStreetMap (OSM) data bring up new issues to consider, including variations of contribution patterns among regions caused by the "digital divide" between developing and developed countries (Goodchild 2007). As of October 2020, the OSM database in Europe was 22.1 GB but only 4.0 GB for the whole continent of Africa (http://download.geofabrik.de). Another aspect is its non-exhaustive nature. Users are more likely to contribute more to one specific place than to others due to, for example, familiarity and pride of place (Forget *et al.* 2018). Finally, there is the issue of quality, which is questionable concerning the contributors' trustworthiness.

Open Government Data (OGD) is another initiative that has been picked up on worldwide, including in countries in developing regions (Vetrò *et al.* 2016). A study of seven African countries (Ghana, Sierra Leone, Tunisia, Morocco, South Africa, Kenya, and Tanzania) showed that by 2017, OGD Web portals had about 1500 data sets in total that were up to date and freely accessible online (Afful-Dadzie and Afful-Dadzie 2017; Lyimo *et al.* 2020). Unlike OSM, government data are authoritative and usually assumed to be of better quality (Fogliaroni *et al.* 2018). OGD data have countrywide coverage and are less affected by the drawbacks related to OSM data discussed in the previous paragraph. The downside of OGD is that publicly available data are usually the product of a derived data model (Shao and Li 2011; Lyimo *et al.* 2020). Hence, volunteered geographic information (such as OSM) and OGD have different modeling schemes and quality characteristics.

When presented with several free and open data sets, selecting the best data set to serve as the training set data becomes important. The selection of sample data points is an essential part of the supervised classification of remotely sensed imagery. The training data set's quality is the key to the accuracy of classification results because inappropriate training samples are the primary source of classification errors (Pal and Mather 2006; Radoux *et al.* 2014; Shao *et al.* 2014). A study by Foody and Arora (1997) demonstrated that the choice of training samples significantly affects the classification results more than does changing the classifier model.

While most research has focused on quality measures for remotely sensed images, little work has focused on quality metrics for training data sets. Ge *et al.* (2008) proposed using rough set theory to analyze sample quality reliability for image classification problems. However, challenges related to selecting a discretization method for the decision table affect this method's adaptation in a broader context. Another related work used open data Portuguese land cover Map (COS) to

Neema Nicodemus Lyimo is with the State Key Laboratory of Information Engineering in Surveying, Mapping, and Remote Sensing, Wuhan University, Wuhan 430079, China; and the Department of Mathematics, Informatics, and Computational Sciences, Solomon Mahlangu College of Science and Education, Sokoine University of Agriculture, PO Box 3038 Morogoro, Tanzania (neemanico@whu.edu.cn).

Fang Luo is with the Wuhan University of Technology, Wuhan, China.

Qimin Cheng is with the Huazhong University of Science and Technology, Wuhan, China.

Hao Peng is with the State Key Laboratory of Information Engineering in Surveying, Mapping, and Remote Sensing, Wuhan University, Wuhan, China.

Contributed by Zhenfeng Shao, October 27, 2020. (sent for review November 23, 2020; reviewed by Md. Enamul Huq, Yaw Danquah Twumasi Nana)

Photogrammetric Engineering & Remote Sensing
Vol. 87, No. 5, May 2021, pp. 339–348.
0099-1112/21/339–348

doi: 10.14358/PERS.87.5.339

Table 1. Environmental and demographic characteristics of Dodoma and Arusha.

City	Climate	Population
Dodoma	Semi-arid	Total population of 410,956 residents, according to the 2012 census (National Bureau of Statistics, Tanzania 2013).
Arusha	Temperate climate	Population of 416,442 in the city center plus 323,198 in the surrounding Arusha districts (National Bureau of Statistics, Tanzania 2013)

Table 2. Product identifiers and acquisition dates of each Landsat scene.

City	Landsat Product Identifier	Acquisition Date	Size of AOI
Dodoma	LC08_L1TP_168064_20190929_20191017_01_T1	29 September 2019	51 × 47 km^2
Arusha	LC08_L1TP_168062_20190929_20191017_01_T1	29 September 2019	14 × 13 km^2

generate training samples for random forest classifiers (Viana 2019). This study explored the k-means clustering technique to select the most representative training samples. These works assessed the quality of training data for image classification tasks from a single type of source or similar data sets. There is a research gap for assessing training data extracted from multiple heterogeneous data sets, particularly open-source–based training data sets.

Open-source data sets vary in many ways, such as in quality, quantity, and modeling schemes, to name a few. Assessment of training sets collected from different sources requires a robust procedure to assess the quality of samples and to optimize them for the classification task. In pixel classification, spectral analysis is the foundation of quality assessment. Various distance functions have been proposed for image quality evaluations. To identify a suitable distance function for accurately processing remotely sensed images, Deborah *et al.* (2015) compared existing distance functions, such as Manhattan, Chebyshev, and Euclidean distance functions; the spectral angle mapper; and the Levenshtein distance. Based on their results, they concluded that Euclidean spectral distance (ESD) is the most appropriate distance function. A study by Forget *et al.* (2018) applied the ESD function to assess the quality of training samples extracted from a single type of open source, namely, OSM, but the study did not discuss the efficiency of its evaluations. This article extends this technique to define the fitness of open-source training samples from volunteered geographic information and OGD. The study relies on the basic ESD principles to model different aspects of data quality in various contexts to effectively measure the quality of training data sets for image classification tasks. The proposed quality measures are evaluated on eight different data sets from two case study cities using Landsat 8 imagery.

Case Studies and Data

Case Studies

Dodoma and Arusha are two cities in Tanzania with different characteristics chosen as study sites (Table 1). Arusha is a major city with a temperate climate. It is a vibrant city that is considered an international diplomatic hub. In 2018, it was declared the capital of the East African Federation. On the other hand, tourism contributes a significant part to Arusha's economy, making it Tanzania's "safari capital" (Bigurube 2004).

In contrast to Arusha, Dodoma, the capital of Tanzania, is a growing city. It has been growing at a slower pace due to the delayed relocation of government activities. This city has a semi-arid climate (Shemsanga *et al.* 2016).

Data

Satellite Imagery

This study used Landsat 8 imagery from the US Geological Survey through the Earth Explorer website (https://earthexplorer.usgs.gov). The scenes were acquired as level 1 data products. Therefore, they are expected to be radiometrically calibrated and orthorectified (Forget *et al.* 2018; Shao *et al.* 2018, 2019; Twumasi *et al.* 2019). Table 2 shows the product identifiers and the acquisition dates of each scene. The latest set of cloud-free scenes were found for both study areas. Both scenes were acquired on the same date. For comparison purposes, it was found necessary to convert the DN values to surface reflectance values. The scene for each city was resized according to the area of interest (AOI) to reduce processing time.

Built-Up Training and Validation Data

OSM provides the most mature and reliable crowdsourced data in this region. The current literature shows that temporal accuracy, up-to-datedness, and lineage quality parameters of OSM in the Tanzania data sets are of higher quality in cities than in peripheral areas (Minghini *et al.* 2018). Therefore, we acquired OSM data for Dodoma and Arusha for August 2020 (Table 3). Data were downloaded via TurboPass. Building footprint layers were downloaded; other data objects were less represented and contained very little information, not sufficient to be used as training data.

Table 3. Data.

Class/ Data Source	Arusha	Dodoma	Source
OSM buildings footprints (OSM BF)	36 236	23 230	overpass-turbo.eu
Distribution points of water users (WDP) data sets	503	412	opendata.go.tz
School facilities data points (SF)	253	156	opendata.go.tz
Health facilities data points (HF)	41	77	opendata.go.tz

Tanzania's OGD source is a rich collection that remains mostly untouched in geospatial applications (Lyimo *et al.* 2020). This collection contains spatial and nonspatial data. In this research, we were interested in spatial data that fall on the built-up area; hence, health facilities (HF), distribution points of water users (WDP), and school facilities (SF) data were selected (Table 3). The assumption behind the inclusion of water users' distribution points is that they represent domestic users' residential locations.

For pre-classification quality assessment, we used no more than 10% of OSM data to reduce processing time; the data were randomly selected across the entire region. However, pre-classification analysis for OGD included complete data sets since they are small in size.

Approximately 2900 polygons were digitized from very high spatial resolution imagery from Google Earth. The data were randomly collected throughout the entire area population to provide a standard measure for comparison of accuracy assessment. Forty percent of the samples (randomly selected) assessed the open data set fitness as training data for a supervised classification task. The remaining amount (60%) of the data set was used to evaluate the performance of built-up classification results.

Training and Validation Samples for Other Land Use/Land Cover Classes

This article's main goal is to assess the quality of open data as training samples; however, none of the sources listed in Table 3 had sufficient data to represent other land cover classes apart from the built-up class in both case study areas. Even if a study has a particular class of interest, conventional supervised classification requires that all categories that occur in the study area be included in the training stage to avoid substantial errors that may be difficult to detect even during accuracy assessment (Foody 2002; Foody *et al.* 2006). Therefore, we collected data from very high spatial resolution imagery from Google Earth for other land cover subclasses in each city, including water, farmland, bare land, vegetation fields, forest/trees, shrubs, and wetlands. After classification, these subclasses were combined into major classes, including water, bare land/low vegetation, and high vegetation.

Methodology

This methodology's basic idea is to generate suitable quality assessment measures by considering the selected classifier's requirements and by assessing variations in characteristics of training samples collected from different open data sources. Figure 1 provides an overview of the proposed methodology for quality assessment and validation.

Selection of Classifier

Several classification algorithms exist. We selected the maximum likelihood classifier (MLC), a simple, common classification algorithm that fits our the purpose, image characteristics, and training data of our analysis. MLC is a pixel-based classification approach that is based on Bayes' theorem. It uses a discriminant function to assign pixels to the class with the highest probability (Ahmad and Quegan, 2012). It achieves that by calculating statistical distances based on the clusters' means and covariance matrices (Ahmad and Quegan 2012; Stein and Tolpekin 2012). MLC is a supervised classification scheme that assumes that spectral classes are statistically characterized by their means and variances (Richards 1993). Statistical distances are probability values that measure spectral uncertainty, and a cell is assigned to the class (cluster) for which it has the lowest uncertainty.

Pre-Classification Training Data Quality Measures

Measuring Spectral Similarity

In pixel-based classification, individual image pixels are characterized by spectral information. Spectral classes represent surface characteristics, or land cover classes. The classification procedure considers a distance to the class's mean as a key to deciding to which class to assign pixels; therefore, assessing spectral differences or similarities is important. We are evaluating the similarity of open-source training data sets

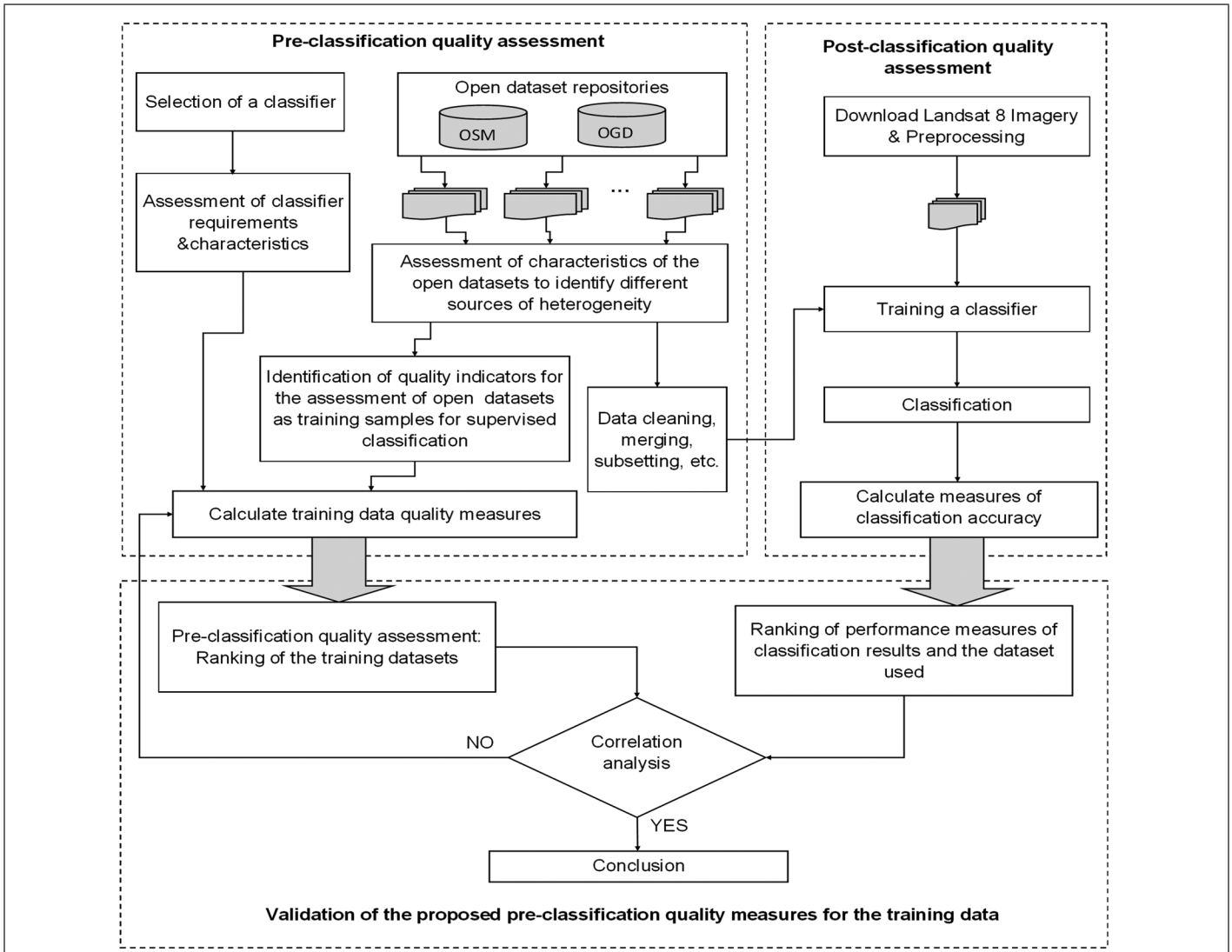

Figure 1. An overview of the methodology.

with reference data. This section improves the ESD formula to determine similarity uncertainties for multiple heterogeneous training data sets. Let us consider k to be one of the heterogeneous training data sets in a given study area and R a reference data set for the study area under investigation. We measure each data set's spectral signature S for the feature class object it represents in six nonthermal Landsat bands,

$$S_k = (\bar{x}_1, \ldots, \bar{x}_n, \ldots, \bar{x}_6), \tag{1}$$

$$S_R = (\bar{y}_1, \ldots, \bar{y}_n, \ldots, \bar{y}_6), \tag{2}$$

where S_k represents the spectral signature of the k^{th} data set with $k = (1, 2, \ldots, m)$, $S_R$represents a spectral signature of a reference data set, and $\bar{x}_n$ and $\bar{y}_n$ are mean pixel values of the featured objects in the two data sets for band n.

According to Forget *et al.* (2018), the ESD d between two featured objects x and y is given by

$$d(x,y) = \sqrt{\sum_{n=1}^{N=6} (\bar{x}_n - \bar{y}_n)^2}. \tag{3}$$

Since we are assessing each of the heterogeneous training data sets separately, Equation 3 can be rewritten to represent the cumulative ESD of data set k:

$$d_k(x,y) = \sqrt{\sum_{n=1}^{N=6} (\bar{x}_{nk} - \bar{y}_n)^2}. \tag{4}$$

We can then normalize the ESD (d_k) results between values a and b for each data set using the normalization Equation 5 to measure the basic similarity between the open-source training data sets and the reference data set. Therefore, we refer to the normalize d_k values as measures of similarity uncertainty (SimU). The smaller the SimU value, the higher the similarity of data set k to the reference data; hence, it is a higher-quality data set and vice versa:

$$\text{SimU} = (b-a) \times \left[\frac{d_k - \min(d_k)}{\max(d_k) - \min(d_k)} \right] + a. \tag{5}$$

Distribution of Training Data Sets in the Feature Space

In this section, we assess the ESD in a different context to determine the intensity of the distribution of feature points in the feature space. Ideally, it is considered that each data set's pixel values will accumulate around certain areas in the feature space and form very dense clusters. The concentration of heterogeneous samples in the clusters will vary from one training data set to another. We assume that a data set whose feature points are very close to one another will have lower uncertainty than a data set whose feature points are far from each other (Figure 2). In other words, a data set whose feature points are very close will have higher quality than a data set whose feature points are far from one another.

This section modifies another Euclidean distance–based method proposed by a recent study of Zhang *et al.* (2019) for modeling uncertainties in remotely sensed images to facilitate uncertainty measurements of heterogeneous training samples in the feature space. The scholars applied the Euclidean distance formula to calculate the distance between feature points in the feature space, as shown in Figure 2. Our feature space is composed of N nonthermal spectral bands of Landsat imagery. We measure the intensity of distribution of the feature points of a given data set in the feature space as follows:

$$\Phi_k = \frac{1}{m} \sum_{j=1}^{m} d_{pj}, \tag{6}$$

where Φ_k represents distribution density of training data set points k, m is the total number of pixels representing the feature points of a training data set k, and d_{pj} represents the ESD of the jth feature point of a training data set to the cluster's reference center point p. We calculate d_{pj} using the following equation:

$$d_{pj} = \sqrt{\sum_{n=1}^{N} \left(f_p^n - f_j^n\right)^2}, \tag{7}$$

where f_p^n represents the nth feature of the central reference feature point p in the cluster and f_j^n is the nth feature of the j^{th} pixel of the training data point in the feature space; in this study, n corresponds to the six spectral nonthermal bands of a multispectral image, and the total number of dimension of feature space N is 6.

To obtain a measure for the assessment of feature space uncertainty (FSU) of the training data sets, we normalize the distribution density Φ_k to values between a and b for each training data set k with Equation 8, and just like Zhang *et al.* (2019), we refer to this type of uncertainty as FSU:

$$\text{FSU} = (b-a) \times \left[\frac{\Phi_k - \min(\Phi_k)}{\max(\Phi_k) - \min(\Phi_k)} \right] + a. \tag{8}$$

Integrating Spectral Similarity and Distribution Measures

We have derived two variations of ESD-based measures from two different contexts or domains. To obtain a more comprehensive measurement model, we combine the two measures using a simple average formula. We refer to the resulting quality measure in Equation 9 as spectral uncertainty (SpU):

$$\text{SpU} = \frac{\text{SimU} + \text{FSU}}{2}. \tag{9}$$

Our analysis evaluates the effectiveness of these measures for ranking the quality of heterogeneous open-source training

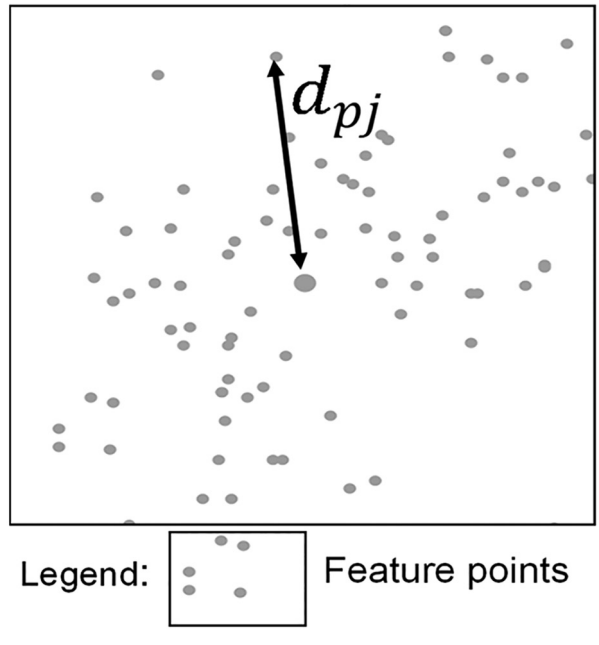

Figure 2. Example of calculating d_{pj}.

data sets for classification tasks. The lower the uncertainty value, the higher the data set quality. The decision to include or exclude a particular data set as a training sample relies on the selected threshold. The assumption is that the lower the uncertainties, the lower the risk of using a low-quality training dataset and hence the higher the classification accuracy.

Classification System

The classification system was developed to represent major land use/land cover types based on the land surface's heterogeneity, as shown in Table 4. The majority filtering process was applied to remove isolated unclassified pixels from the classification output. Majority analysis filtering is a standard smoothing procedure to reduce some salt-and-pepper noises (Su 2016).

Table 4. Land use/land cover classification system.

Land Use/Land Cover	Description
Built-up	Residential/industrial/commercial areas where rooftops dominate
Bare land/low vegetation	Cleared land/farmland/bare land/areas with low vegetation growing
High vegetation	Areas covered with trees/shrubs/forest
Water	Water bodies, such as reservoirs, ponds, and rivers

Post-Classification Accuracy Measures

Even though pre-classification quality assessment measures provide a useful prediction of how the training samples will perform, the real effect of the given training data's quality will be observed in the classification results (Ge *et al.* 2012). Therefore, we assess the correlation between pre-classification quality measures and post-classification quality measures. For comparison purposes, we evaluate the classification results of a given city with the same test set.

Here we consider accuracy measures, which allow us to effectively compare probabilities of either correct or incorrect classification for each result based on the training data set. The first measure is overall accuracy, referred to as a proportion of correctly classified pixels, given as (Foody 2002)

$$P_C = \sum_{k=1}^{q} P_{kk}. \quad (10)$$

Another measure is the kappa coefficient. Kappa statistics are useful for evaluation and comparison classification results based on different data or methods (Shao *et al.* 2017). Let x_{ij} denote the element of the error matrix in row i and column j, r denote the number of classes, and N denote the total sum of all elements of the error matrix. Then kappa coefficient k is computed as (Cohen 1960)

$$k = \frac{N\sum_{i=1}^{r} x_{ii} - \sum_{i=1}^{r} x_{i+}x_{+i}}{N^2 - \sum_{i=1}^{r} x_{i+}x_{+i}}, \quad (11)$$

where $x_{i+} = \sum_{j=1}^{r} x_{ij}$ and $x_{+i} = \sum_{j=1}^{r} x_{ji}$ are the sums of all elements in row i and column i, respectively.

Other measures include errors of omission and commission and user and producer errors. Omission errors refer to the number of pixels that were not included in interpreting the class results. In contrast, commission errors occur when samples have been wrongly classified as belonging to a particular class (Stehman and Czaplewski 1998; Stein and Tolpekin 2012; Sumari *et al.* 2020).

Analysis and Results

The study has derived quality measures based on ESD metrics to assess the fitness of four open-source training data sets for urban area classification in two cities. The open data are from two sources: OSM and OGD. We observe variations among the open-source data sets. For example, in OSM, we obtained building footprints (OSM BF) that contain 20 000 to 30 000 polygons. In OGD, we found three data sets that were related to built-up land use/land cover: SF, HF, and WDP. The amount of OGD data sets varied from hundreds to a few tenths. We also observed that different open data sources have variations in data types and formats. Data set types are of two categories: points and polygons. The data sets vary in quality, data size, and data type; considering all these variations, we would like to observe whether the proposed quality measures can successfully rank the quality of the training data sets for image classification.

Apart from built-up related data, we did not find enough freely available data for an accurate representation of other land use/land cover classes. Since we did not find such data in both study areas, we collected data from Google Earth for other land cover classes for each city. Training samples extracted from open sources were used for the classification of the built-up class. Therefore, quality assessments are carried out based on the variations of several open data sets used to train the built-up class.

Pre-Classification Quality Assessments

In the methodology, we derived three quality measures based on the ESD function: SimU, FSU, and a combination of the two, referred to as SpU. SimU measures the similarity between open-source training data sets and reference data using the mean of the ESD. FSU is a normalized measure of the spread or the concentration of the feature points of the data sets in the feature space. Finally, SpU takes advantage of the two measures' varied capabilities by combining them to determine the overall spectral uncertainty of the training data sets.

Tables 5 and 6 show a summary of pre-classification quality assessments. A lower d_k leads to a lower SimU value; hence, the data set is ranked as having a higher similarity in quality with the reference data set since it has lower similarity uncertainties and vice versa. On the other hand, when a data set cluster in the feature space is densely concentrated (Φ_k), it leads to lower FSU, reflecting that the data set has good spectral coverage to represent a given training class. Values of 0.1 and 0.9 were used as scale values a and b in Equations 5 and 8.

The graphs in Figures 3 and 4 provide us with some visualizations of the variations in the data sets. We observe limited variations of spectral values for the data sets in Arusha compared to Dodoma. In Dodoma, the data are more similar in lower bands, but we notice a higher discrepancy between bands 5 and 7, except for OSM BF and reference data.

The WDP data set in Figure 3 has the largest difference from reference data in bands 6 and 7. Also, in Table 5, the WDP has a d_k value of 0.099329; the difference is about seven times larger than OSM BF data. For clarity, we analyzed this section further in the two-dimensional feature space in Figure 5. The results show a significant shift of WDP data point for bands 6 and 7 in Dodoma.

The box chart type in Figure 5 enables us to picture and compare the distribution of the data sets by grouping them based on five fundamental values: minimum, first quartile, median, third quartile, and maximum. The chart's box section is also referred to as the interquartile range (IQR); it represents 50% of the data values. The graph also shows the minimum and maximum spectral reflectance values in the data set via vertical statistical lines extending from the box.

OSM has the largest data sets in both cities, while HF data sets are the smallest, with the least having 41 data points

Table 5. Summary of pre-classification quality assessments for Dodoma.

	OSM BF Data	SF Data	WDP Data	HF Data
Mean spectral distance (d_k)	0.014 779	0.05 442 085	0.099 329	0.047 625
Similarity uncertainty (SimU)	0.1	0.475 088 969	0.9	0.410 788
Distribution of spectral values in the feature space (Φ_k)	7.16E-07	0.000 355 692	0.000 242	0.001 162
Feature space uncertainty (FSU)	0.1	0.344 628 671	0.266 462	0.9
Spectral uncertainty (SpU)	0.1	0.40 985 882	0.583 231	0.655 394

Table 6. Summary of pre-classification quality assessments for Arusha.

	SF Data	OSM BF Data	HF Data	WDP Data
Mean spectral distance (d_k)	0.0 111 979	0.031 139 503	0.012 942	0.028 260 712
Similarity uncertainty (SimU)	0.1	0.9	0.169 967	0.78 451 115
Distribution of spectral values in the feature space (Φ_k)	4.57 057E-05	1.16 444E-06	0.000 168	5.72 079E-05
Feature Space uncertainty (FSU)	0.313 482 577	0.1	0.9	0.36 861 173
Spectral uncertainty (SpU)	0.206 741 288	0.5	0.534 983	0.57 656 144

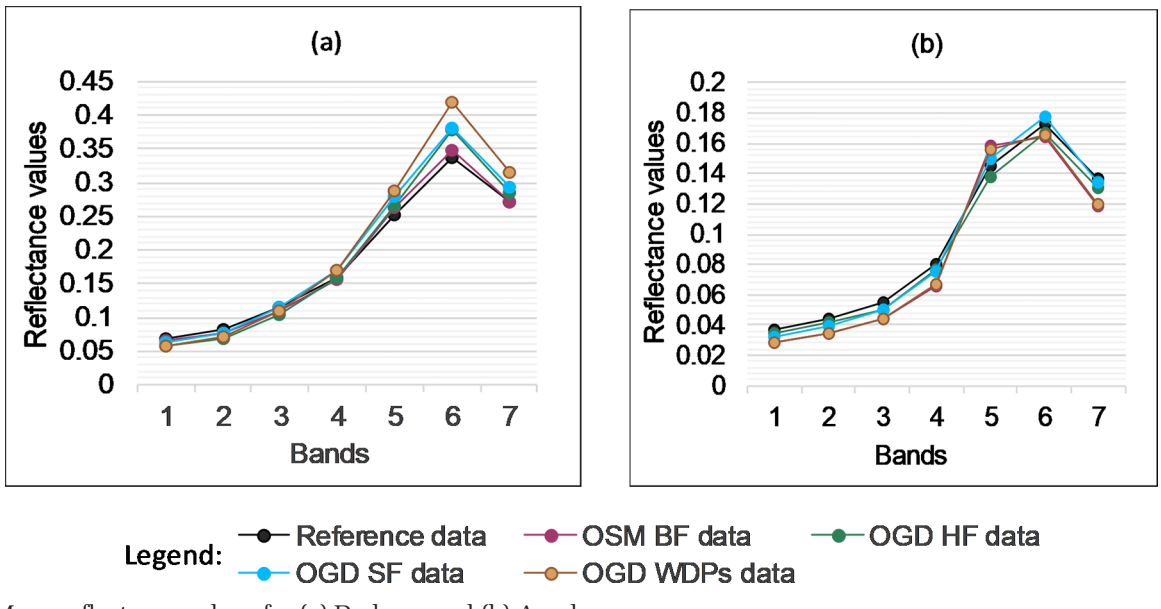

Figure 3. Mean reflectance values for (a) Dodoma and (b) Arusha.

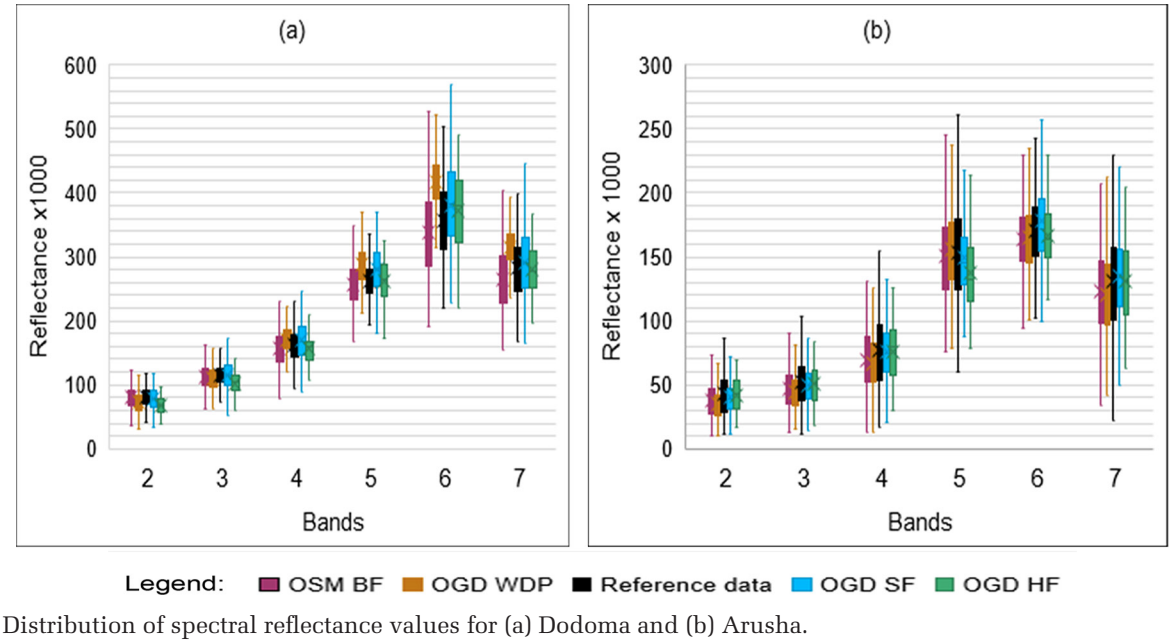

Figure 4. Distribution of spectral reflectance values for (a) Dodoma and (b) Arusha.

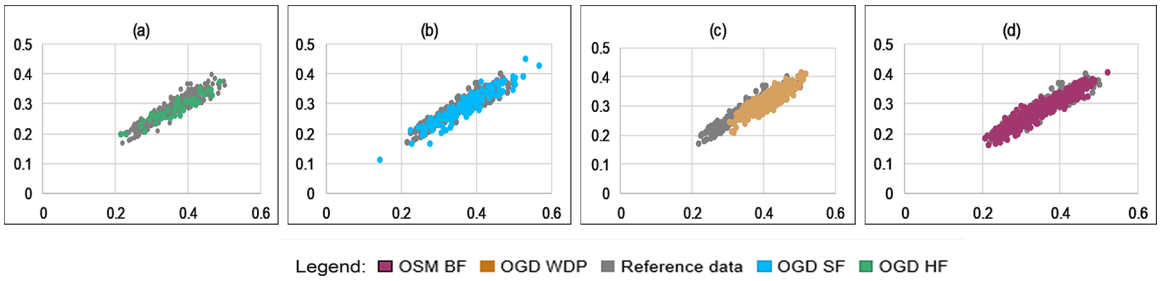

Figure 5. Distribution the open data sets in a two-dimensional feature space of bands 6 and 7 for Dodoma.

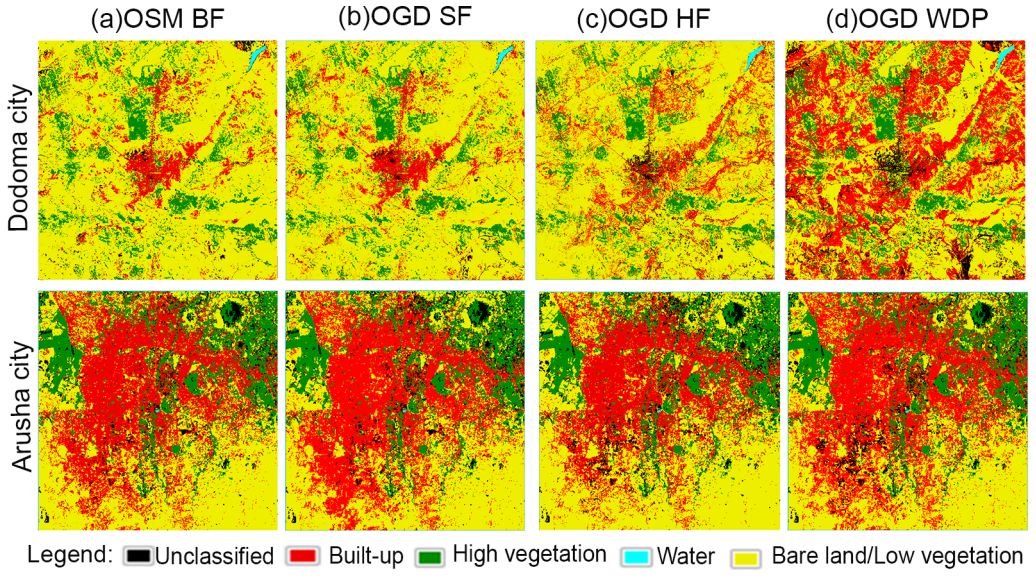

Figure 6. Classification results for Dodomaand Arusha.

available only in Dodoma. However, we can observe that some data sets, such as SF, have comparable distribution coverage despite having a small size compared to OSM.

Classification Results

Classification accuracy measures were derived from the error matrix of classified images. The proposed post-classification measures include overall accuracy (OA, with all classes), overall kappa coefficient including all classes (OKC), built-up class accuracy (BA), kappa coefficient for built-up class (KCB), built-up unclassified pixels, a built-up area classified as bare land, a built-up area classified as an area with thick vegetation, and the errors of commission and omission for built-up class. Results of the OA, which include the OSM BF, OGD SF, OGD HF, and OGD WDP data sets, are 86.55%, 85.01%, 71.65%, and 59.44% in Dodoma and 85.86%, 86.52%, 84.42%, and 81.09% in Arusha, respectively (Figure 6).

Discussion

Role of the ESD for Modeling Quality Measures

The modeling quality measures depends on the assessment of spectral information in various contexts. Inspired by ESD metrics, we derived three quality measures for open-source training data sets: SimU, FSU, and SpU. Tables 7 and 8 present a visualization of a ranking correlation between pre-classification and post-classification quality measures.

A total of eight data sets were collected from OSM and OGD. The data sets are truly heterogeneous, with significant variations from one another in terms of quality, quantity, and data type. The evaluations are focused on the built-up class to determine the variations of different open data sets used for the training.

Tables 7 and 8 show a comparison between the proposed quality measures and overall built-up class accuracy. FSU ranks only five out of eight cases correctly, while SimU ranks six out of eight tested cases. When used alone, none of the two measures rank the quality of all the data sets successfully; each one is limited since it relies on a single perspective to assess the data. But when combined to produce a third metric, SpU, a more comprehensive quality assessment happens to rank all eight cases correctly. SpU performance proves the versatility of ESD for modeling different aspects of data to facilitate measuring the quality of training data sets for image classification purposes, especially those collected from diverse free sources.

Previous works have considered a single perspective for assessing data quality for training purposes; moreover, these works use data from single or similar sources (Forget *et al.* 2018; Viana 2019). Our results demonstrate that modeling the quality of training data from diverse open sources is more complex and takes a combination of different models for the successful quantification of quality. Spectral analysis is the key to examining data set quality for pixel-based classification. ESD has proven to be a powerful metric for modeling different aspects of quality measures on spectral similarity and FSU.

However, it should be noted that each data set should meet a minimum limit of data size for the selected classifier for a reliable SpU ranking. For example, MLC requires at least 10 p, where p is the number of bands used; for six bands, a data set with fewer than 60 pixels may have some inconsistencies, like in the case of HF data sets with 41 pixels for Dodoma. Its classification results have some discrepancies with kappa accuracy. However, some researchers in accuracy assessment have cautioned against using the kappa coefficient (Stehman

Table 7. Comparison between pre-classification and post-classification quality measures for Dodoma.

	OSM BF Data	SF Data	WDP Data	HF Data
Data set properties				
Data type	Polygons	Points	Points	Points
Data size (in pixel counts)	20 635	153	410	41
Quality measures				
Similarity uncertainty (SimU)	0.1	0.47 509	0.9	0.410 788
Feature space uncertainty (FSU)	0.1	0.34 463	0.266 462	0.9
Spectral uncertainty (SpU)	0.1	0.40 986	0.5 832 308	0.655 393
Post-classification accuracy assessment				
Built-up class accuracy (%)	80.64	77.81	41.48	37.13
Kappa coefficient for built-up class	0.73	0.70	0.12	0.31
Built-up area classified as bare land (%)	12.33	11.75	16.03	30.24
Built-up area classified as thick vegetation (%)	0.15	0.22	0.29	0.15
Built-up area classified as area water (%)	0	0	0	0
Error of commission with respect to built-up class (%)	15.76	17.78	56.03	32.01
Error of omission with respect to built-up class (%)	19.36	22.19	58.52	62.87
Overall accuracy including all classes (%)	86.55	85.01	59.44	71.65
Kappa coefficient including all classes	0.7906	0.77	0.43	0.57

Table 8. Comparison between pre-classification and post-classification quality measures for Arusha.

	SF Data	OSM BF Data	HF Data	WDP Data
Data set properties				
Data type	Points	Polygons	Points	Points
Data size (in pixel counts)	245	26 742	77	494
Quality measures				
Similarity uncertainty (SimU)	0.1	0.9	0.169 967	0.7 845 112
Feature space uncertainty (FSU)	0.31 348	0.1	0.9	0.3 686 117
Spectral uncertainty (SpU)	0.20 674	0.5	0.534 983	0.5 765 614
Post-classification accuracy assessment				
Built-up class accuracy (%)	84.29	83.10	78.68	75.62
Kappa coefficient for built-up class	0.78	0.765	0.751	0.673
Built-up area classified as bare land (%)	7.69	9.75	11.39	11.57
Built-up area classified as thick vegetation (%)	3.25	3.14	0.32	3.01
Built-up area classified as area water (%)	0	0	0	0
Error of commission with respect to built-up class (%)	2.93	2.82	3.77	3.82
Error of omission with respect to built-up class (%)	15.71	16.90	21.32	24.38
Overall accuracy including all classes (%)	86.52	85.86	84.42	81.09
Kappa coefficient including all classes	0.7741	0.764	0.7638	0.6963

and Czaplewski 1998); our future works will address this aspect further.

Effect of Different Sizes of Training Sets

In some cases, the increase in training data size has been shown to influence the decrease in FSU and even to minimize the overall SpU, leading to better classification results despite having a higher SimU. For example, Arusha OSM BF data with the largest SimU are ranked as the second-best data set by SpU. Compared with post-classification accuracy, it is also ranked as the second-best results in that city. It should be noted that OSM BF is the largest training data set in this city and so has smaller FSU. On the other hand, results show that when the data size is below a minimum requirement, significant increases occur in FSU, and overall performance drops drastically, such as in the case of classification results based on HF data points (37.13%). Despite the HF training class having a smaller SimU than SF data in Dodoma, the HF class size is almost half of the SF data set. Therefore, we can conclude that when all data sets are within the acceptable range of SimU, the amount of data size can positively influence the results. The derived measures can capture this aspect before classification begins and rank the data sets accordingly.

Impact of Different Data Types

The open-source training data sets comprise two main data types: points and polygons. For spectrally homogeneous land cover class, point-based training data sets have similar

performance to polygon-based training data (Chen and Stow 2002). For spectrally heterogeneous surface cover, polygons have the advantage of capturing a better range of available spectra information. However, pixels found along the borders of polygons usually contain mixed spectral values and can be detrimental to classification accuracy (Boudewyn *et al.* 2000). This phenomenon may impact the classification results of images with higher resolution than coarse resolution levels (Chen and Stow 2002). This factor's influence cannot be noticed for small buildings whose area is less than or equal to a pixel for medium-resolution imagery like *Landsat 8*. However, it was interesting to discover that there are approximately 266 polygons whose area is larger than one or more pixels of *Landsat 8* in Arusha. Despite the possible effect of mixed pixels in the boundaries, these polygons also have a higher chance of capturing pixels of higher purity levels. Hence, the polygons ensure more comprehensive coverage of the spectral range for a given training class.

This phenomenon is well modeled with FSU quality measures. Consider OSM BF and WDP results for Arusha. OSM BF is ranked with the highest level of uncertainty regarding SimU (0.9) and hence is considered a data set that is more dissimilar to the reference data set than WDP (SimU ≈0.78). However, this data set has the lowest FSU (0.1); a plausible explanation for this is the contribution of good spectral coverage caused by the polygon data type in this set. On the other hand, a reason for a lower SimU would be due to some mixed pixels of polygon boundaries, but it seems that this was not big enough to affect the overall quality of the data in SpU (0.5). Furthermore, point-based training data are more likely to have gaps in spectral information, causing a higher FSU and leading to an increase of unclassified pixels, as in the HF results in both case study areas.

Conclusions

Existing research has assessed training samples from a single or from similar open data sources. Such procedures may not be sufficient to examine the quality of heterogeneous data sets extracted from diverse sources. Eight different data sets from OSM and OGD were collected. We relied on ESD metrics to derive three quality measures for assessing the quality of the open-source training samples: SimU, FSU, and SpU. The lower the uncertainties, the higher the data quality. The study analyzes the relationship between pre-classification quality measures and post-classification accuracy measures.

Correlation analysis aimed to determine the robustness of the proposed quality measures against data sets of varying quality, data set size, and data types. A comparison with classification accuracy proves that SpU, which is the combination of FSU and SimU, can successfully rank the quality of training data sets despite their variations. These findings demonstrate the versatility of ESD for modeling different quality aspects of heterogeneous data sets based on spectral information in various contexts.

Other key findings include the following:

- Polygon data set type can help to reduce FSU because of continuous coverage of spectral information; however, at the same time, polygon boundaries can be the primary source of mixed pixels, which increases uncertainties in similarity measurements.
- Point data type can lead to higher FSU because of limited coverage of spectral information, especially for heterogeneous land surface covers, such as in urban areas.
- An increase in data set size has a positive influence and can even minimize the overall spectral uncertainty for data sets with low similarity uncertainties.
- However, data set size below the minimum requirements for a selected classifier can cause inconsistencies in the results.

This article provides a more comprehensive approach for the quality assessment of open-source training sets than existing works. With the increasing availability of open data, this method facilitates predetermination of the data set's quality to empower researchers to choose suitable training sets for image classification. It also allows optimizing the data sets before the classification procedures take place.

This work has applied a simple average to combine two derived quality evaluation metrics to measure the spectral uncertainty of heterogeneous data sets in different contexts. Our future works will include experiments with different weights and combinations of parameters to further test and analyze the validity, applicability, and robustness of the proposed approach and to optimize it accordingly.

Acknowledgments

This work was supported by the National Key Research and Development Program of China (2018YFB0505401), the Research Project from the Ministry of Natural Resources of China under grant 4201-240100123, the National Natural Science Foundation of China under grants 41771452, 41771454, 41890820, and 41901340; the Natural Science Fund of Hubei Province in China under grant 2018CFA007; and National Major Project on High Resolution Earth Observation System (GFZX0403260306). The authors are sincerely grateful to the editors and anonymous reviewers for their valuable suggestions and comments that helped us improve this article significantly.

References

Afful-Dadzie, E., and A. Afful-Dadzie. 2017. Open Government Data in Africa: A preference elicitation analysis of media practitioners. *Government Information Quarterly* 34 (2):244–255. <https://doi.org/10.1016/j.giq.2017.02.005>

Ahmad, A., and S. Quegan. 2012. Analysis of maximum likelihood classification on multispectral data. *Applied Mathematical Sciences* 6 (129–132):6425–6436.

Bielecka, E., and E. Burek. 2019. Spatial data quality and uncertainty publication patterns and trends by bibliometric analysis. *GeoScape* 11 (1):219–235. <https://doi.org/10.1515/geo-2019-0018>

Bigurube, G. 2004. Tanzania national parks bronchure. TANAPA website, Tanzania. <https://www.tanzania.go.tz/egov_uploads/documents/tanapa_sw.pdf> Accessed 6 December 2020.

Boudewyn, P., D. Seemann, M. Wulder and S. Magnussen. 2000. The effects of polygon boundary pixels on image classification accuracy. Pages 637–643 in *Remote Sensing and Spatial Data Integration: Measuring, Monitoring and Modeling, 22nd Symposium of the Canadian Remote Sensing Society*, held in XXXX, XXXX. XXXX: XXXX.

Chen, D. M., and D. Stow. 2002. The effect of training strategies on supervised classification at different spatial resolutions. *Photogrammetric Engineering and Remote Sensing* 68 (11):1155–1161.

Cohen, J. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement* 20 (1):37–46.

Deborah, H., N. Richard and J. Y. Hardeberg. 2015. A Comprehensive evaluation of spectral distance functions and metrics for hyperspectral image processing. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 8 (6):3224–3234. <https://doi.org/10.1109/JSTARS.2015.2403257>

Deren, L., M. Jun, C. Tao, J. L. van Genderen and Z. Shao. 2019. Challenges and opportunities for the development of MEGACITIES. *International Journal of Digital Earth* 12 (12):1382–1395. <https://doi.org/10.1080/17538947.2018.1512662>

Deren, L., Y. Yao, Z. Shao and L. Wang. 2014. From digital Earth to smart Earth. *Chinese Science Bulletin* 59 (8):722–733.

Fogliaroni, P., F. D'Antonio and E. Clementini. 2018. Data trustworthiness and user reputation as indicators of VGI quality. *Geo-Spatial Information Science* 21 (3):213–233. <https://doi.org/10.1080/10095020.2018.1496556>

Foody, G. M. 2002. Status of land cover classification accuracy assessment. *Remote Sensing of Environment* 80:185–201.

Foody, G., and M. Arora. 1997. An evaluation of some factors affecting the accuracy of classification by an artificial neural network. *International Journal Remote Sensing* 18:799–810.

Foody, G. M., A. Mathur,C. Sanchez-Hernandez and D. S. Boyd. 2006. Training set size requirements for the classification of a specific class. *Remote Sensing of Environment* 104 (1):1–14. <https://doi.org/10.1016/j.rse.2006.03.004>

Forget, Y., C. Linard and M. Gilbert. 2018. Supervised classification of built-up areas in Sub-Saharan African cities using Landsat imagery and openstreetmap. *Remote Sensing* 10 (7):1–16. <https://doi.org/10.3390/rs10071145>

Ge, Y., H. Bai, S. Li and D. Li. 2008. Exploring the sample quality using rough sets theory for the supervised classification of remotely sensed imagery. *Geo-Spatial Information Science* 11 (2):95–102. <https://doi.org/10.1007/s11806-008-0020-0>

Ge, Y., H. Bai, J. Wang and F. Cao. 2012. Assessing the quality of training data in the supervised classification of remotely sensed imagery: A correlation analysis. *Journal of Spatial Science* 57 (2):135–152. <https://doi.org/10.1080/14498596.2012.733616>

Goodchild, M. F. 2007. Citizens as sensors: The world of volunteered geography. *GeoJournal* 69 (4):211–221. <https://doi.org/10.1007/s10708-007-9111-y>

Lyimo, N. N., Z. Shao, A. M. Ally, N.Y.D. Twumasi, O. Altan and C. A. Sanga. 2020. A fuzzy logic-based approach for modelling uncertainty in open geospatial data on landfill suitability analysis. *ISPRS International Journal of Geo-Information* <https://doi.org/10.3390/ijgi9120737>

Minghini, M., M. A. Brovelli and F. Frassinelli. 2018. An open-source approach for the intrinsic assessment of the temporal accuracy, up-to-dateness, and lineage of OpenStreetMap. *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences—ISPRS Archives* 42 (4W8):147–154. <https://doi.org/10.5194/isprs-archives-XLII-4-W8-147-2018>

National Bureau of Statistics, Tanzania. 2013. Sensa2012. <http://www.dataforall.org/CensusInfoTanzania/libraries/aspx/Home.aspx> Accessed 28 November 2020.

Pal, M., and P. Mather. 2006. Some issues in the classification of DAIS hyperspectral data. *International Journal of Remote Sensing* 27:2895–2916.

Radoux, J., C. Lamarche, E. Van Bogaert, S. Bontemps, C. Brockmann and P. Defourny. 2014. Automated training sample extraction for global land cover mapping. *Remote Sensing* 6 (5):3965–3987. <https://doi.org/10.3390/rs6053965>

Richards, J. A. 1993. *Remote Sensing Digital Image Analysis: An Introduction*. Berlin: Springer-Verlag.

Shao, Z., and J. Cai. 2018. Remote sensing image fusion with deep convolutional neural network. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 11 (5):1656–1669. <https://doi.org/10.1109/JSTARS.2018.2805923>

Shao, Z., J. Cai, P. Fu, L. Hu and T. Liu. 2019. Deep learning-based fusion of Landsat-8 and Sentinel-2 images for a harmonized surface reflectance product. *Remote Sensing of Environment* 235:111425. <https://doi.org/10.1016/j.rse.2019.111425>

Shao, Z., J. Deng, L. Wang, Y. Fan, N. S. Sumari and Q. Cheng. 2017. Fuzzy AutoEncode based cloud detection for remote sensing imagery. *Remote Sensing*. <https://doi.org/10.3390/rs9040311>

Shao, Z., and D. Li . 2011. Image City sharing platform and its typical applications. *Science China Information Sciences* 54:1738–1746. <doi:10.1007/s11432-011-4307-7>

Shao, Z., K. Yang and W. Zhou. 2018. Performance evaluation of single-label and multi-label remote sensing image retrieval using a dense labeling dataset. *Remote Sensing* 10 (6):964. <doi:10.3390/rs10060964>

Shao, Z., L. Zhang, X. Zhou and L. Ding. 2014. A novel hierarchical semisupervised SVM for classification of hyperspectral images. *IEEE Geoscience and Remote Sensing Letters* 11 (9):1609–1613. <doi:10.1109/LGRS.2014.2302034>

Shao, Z., W. Zhou, X. Deng, M. Zhang and Q. Cheng. 2020. Multilabel remote sensing image retrieval based on fully convolutional network. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 13:318–328.

Shemsanga, C., A.N.N. Muzuka, L. Martz, H. Komakech and A. N. Omambia. 2016. Statistics in climate variability, dry spells, and implications for local livelihoods in semiarid regions of Tanzania: The way forward. In *Handbook of Climate Change Mitigation and Adaptation, Second Edition*, vol. 2, 801–848. Berlin: Springer International Publishing. <https://doi.org/10.1007/978-3-319-14409-2_66>

Stanislawski, L. V., B. Buttenfield, P. Bereuter and C. Brewer. 2014. Abstracting geographic information in a data rich world. In *Abstracting Geographic Information in a Data Rich World: Methodologies and Applications of Map Generalisation*, edited by Dirk Burghardt, Cécile Duchêne, and William Mackaness, 1–15. Berlin: Springer International Publishing. <https://doi.org/10.1007/978-3-319-00203-3>

Stehman, S. V., and R. L. Czaplewski. 1998. Design and analysis for thematic map accuracy assessment: Fundamental principles. *Remote Sensing of Environment* 64 (3):331–344. <https://doi.org/10.1016/S0034-4257(98)00010-8>

Stein, A., and V. Tolpekin. 2012. *The Core of GIScience: A System-Based Approach*. Enschede, Netherlands: International Institute for Geo-Information Science and Earth Observation.

Sumari, N. S., P. B. Cobbinah, F. Ujoh and G. Xu. 2020. On the absurdity of rapid urbanization: Spatio-temporal analysis of land-use changes in Morogoro, Tanzania. *Cities*. <https://doi.org/10.1016/j.cities.2020.102876>

Twumasi, N.Y.D., Z. Shao and O. Altan. 2019. Mapping built-up areas using two band ratio on Landsat imagery of Accra in Ghana from 1980 to 2017. *Applied Ecology and Environmental Research* 17 (6):13147–13168. <https://doi.org/10.15666/aeer/1706_1314713168>

Vetrò, A., L. Canova, M. Torchiano, C. O. Minotas, R. Lemma and F. Morando. 2016. Open data quality measurement framework: Definition and application to Open Government Data. *Government Information Quarterly* 33 (2):325–337. <https://doi.org/10.1016/j.giq.2016.02.001>

Viana, C. M. 2019. Training samples from open data for satellite imagery classification: Using K-means clustering algorithm. Pages 17–20 in *Proceedings of the 22nd AGILE Conference on Geo-Information Science*, held in Limassol, Cyprus, XXXX. XXXX: XXXX.

Yin, L., Q. Cheng, Z. Wang, and Z. Shao. 2015. "Big data" for pedestrian volume: Exploring the use of Google Street View images for pedestrian counts. *Applied Geography* 63:337–345.

Zhang, Q., P. Zhang and Y. Xiao. 2019. A modeling and measurement approach for the uncertainty of features extracted from remote sensing images. *Remote Sensing* 11 (16). <https://doi.org/10.3390/rs11161841>

Zhou, W., S. Newsam, C. Li and Z. Shao. 2018. PatternNet: A benchmark dataset for performance evaluation of remote sensing image retrieval. *ISPRS Journal of Photogrammetry and Remote Sensing* 145:197–209.