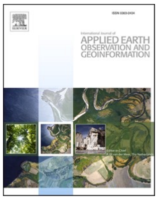
International Journal of Applied Earth Observations and Geoinformation 94 (2021) 102235

Contents lists available at ScienceDirect

International Journal of Applied Earth Observations and Geoinformation

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



Determining representative sample size for validation of continuous, large continental remote sensing data

Megan L. Blatchford *, Chris M. Mannaerts, Yijian Zeng

Department of Water Resources, ITC Faculty of Geo-Information Science and Earth Observation, University of Twente, Hengelostraat 99, 7514 AE, Enschede, the Netherlands

ARTICLE INFO

Keywords:
Progressive sampling
Confidence interval
Maximum entropy probability
Performance measure
Big data

ABSTRACT

The validation of global remote sensing data comprises multiple methods including comparison to field measurements, cross-comparisons and verification of physical consistency. Physical consistency and cross-comparisons are typically assessed for all pixels of the entire product extent, which requires intensive computing. This paper proposes a statistically representative sampling approach to reduce time and efforts associated with validations of remote sensing data having big data volume. A progressive sampling approach, as typically applied in machine learning to train algorithms, combined with two performance measures, was applied to estimate the required sample size. The confidence interval (CI) and maximum entropy probability distribution were used as indicators to represent accuracy. The approach was tested on 8 continental remote sensing-based data products over the Middle East and Africa. Without the consideration of climate classes, a sample size of 10,000–100,000, dependent on the product, met the nominally set CI and entropy indicators. This corresponds to <0.01 % of the total image for the high-resolution images. All continuous datasets showed the same trend of CI and entropy with increasing sample size. The actual evapotranspiration and interception (ETIa) product was further analysed based on climate classes, which increased the sample size required to meet performance requirements, but was still determined to be significantly less than the entire dataset size. The proposed approach can significantly reduce the processing time while still providing a statistically valid representation of a large remote sensing dataset. This can be useful as more high-resolution remote sensing data becomes available.

1. Introduction

Global and continental remote sensing data sets are increasingly available in terms of access and dataset type. Satellite temporal and spatial resolution, i.e. the pixel size and return period, are also increasing as satellite technology improves. As the dataset resolution improves, the processing time increases not only for production but also for validation.

A common part of validation procedures for continental (and larger) datasets is verification of spatial and temporal consistency and cross-comparison to other datasets (Loew et al., 2017; Zeng et al., 2019, 2015). Spatial and temporal consistency is a physical consistency test that considers the variation and relative uncertainty of the product over space and time, while cross-comparison compares the product directly to a reference product developed by a different producer, or using a different satellite, algorithm etc. Physical consistency analyses variation of the product, dependent on factors such as climate and season. This is seen as a component of the validation strategies of various producers including the Copernicus Global Land Surface Products, which includes vegetation (e.g. dry matter productivity and leaf area index), energy (e. g. surface albedo), water (e.g. water bodies) and cryosphere products (Smets et al., 2013), Advanced Along-Track Scanning Radiometer (AATSR) Land Surface Temperature (LST) Validation Strategy (Schneider et al., 2012), the FAO Water Productivity Open-access portal (WaPOR) of Remotely sensed derived data validation methodology, which includes vegetation (e.g. net primary productivity - NPP) and water (precipitation and evapotranspiration - ETa) products (FAO, 2018) and Moderate-resolution Imaging Spectroradiometer (MODIS) land validation strategy (MODLAND), which includes vegetation (e.g. vegetation indices and NPP) and energy (e.g. LST) (Morisette et al., 2002).

Typically, physical consistency and cross-comparisons are evaluated over the entire extent of the dataset on a pixel-by-pixel basis to determine spatial and temporal trends and differences. This not only requires

* Corresponding author.
E-mail address: m.l.blatchford@utwente.nl (M.L. Blatchford).

https://doi.org/10.1016/j.jag.2020.102235
Received 14 April 2020; Received in revised form 3 September 2020; Accepted 3 September 2020
Available online 21 September 2020

substantial computational costs as spatial resolution increases but can also be excessive to understand the performance of the dataset. Alternatively, samples can provide enough insight to accuracy with less computation to evaluate these arbitrarily large datasets. In several fields, including in land use classification (Heydari and Mountrakis, 2018), machine learning, bioinformatics (Kim, 2009), clinical studies (Gupta et al., 2016; Kirby et al., 2002; Lachin, 1981) and classifier design studies (Fukunaga and Hayes, 1989), scaling-down techniques are used to approach the problem of training large datasets by selecting a sample of the data which is meant to accurately represent the entire dataset. However, determining the appropriate sample size with large datasets is not always obvious and has not been applied in validation.

Machine learning frequently deals with developing and training algorithms for large databases. In machine learning the primary categories of scaling-down sampling methods are random selection, active learning techniques and progressive sampling (ElRafey and Wojtusiak, 2017). Random sampling uses passive learning, active learning uses semi-supervised machine learning to choose data from which it learns. Active learning and passive learning methods typically use an arbitrary, predefined sample size (Warmuth et al., 2003). Active learning algorithms seek to select the most informative cases for training while progressive sampling aims to minimize the amount of computation for a given performance target.

Progressive sampling incrementally increases the sample size until the accuracy of the algorithm no longer improves, or converges (Luo, 2016) and is designed to efficiently produce models with high accuracy (Meek et al., 2002). This prevents processing the entire database, which may be resources heavy (Ng and Dash, 2006). Progressive sampling helps balance the prediction accuracy and the data processing effort (Sarkar et al., 2016) to determine the optimal sample size (Gu et al., 2001).

Progressive sampling is sparsely identified in literature in remote sensing applications. It has been applied to learn neural network ensembles of arbitrarily large datasets (Peng et al., 2004), digital terrain modeling (DEM) data acquisition (Chen and Li, 2013; Makarovic, 1973), and has been integrated into digital image mapping (Rauhala, 1989). Most commonly, progressive sampling has been used in clinical studies (Figueroa et al., 2012) and in training algorithms in machine learning and association rules in data mining (Last, 2009; Ng and Dash, 2006; Umarani and Punithavalli, 2011; Zeng and Luo, 2017). Examples in other fields include field sample design for ecological studies (Stein and Ettema, 2003) and argo-ecological characterization (Steduto and Todorovic, 2001).

The accuracy or performance of a product can be estimated through many metrics. For discrete variables, such as land classification or in machine learning applications, the accuracy is often taken as the number of correct predictions over the total number of data predictions (Congalton and Green, 2009; Foody, 2002). For continuous datasets, the accuracy metrics used are more diverse. Regression metrics are commonly used when true value is known. However, physical consistency and cross-comparisons is performed in the absence of known true values and is a method to observe comparative performance of the data over space and time rather than provide absolute accuracy.

The coefficient of variation (CV) is an index of reliability or relative variability, which is commonly used in several fields of science (Payton, 1996; Reed et al., 2002; Schectman, 2013). The confidence interval (CI) is expressed in terms of the variation around the expected value in terms of the CV or standard deviation. The CI can be used to express relative accuracy (Burt et al., 1997; Young and Lewis, 1997). The Principle of Maximum Entropy states that the distribution with the maximum entropy best matches the current state of knowledge and provides a measure of the amount of information needed to represent an outcome from a probability distribution for a random variable. Entropy was first formulated to understand the diversity and uniformity of discrete variables otherwise known as the Shannon's Index (Shannon, 1948). It was later generalised to the differential entropy and to continuous random variables (Jaynes, 1957; Santamaría-Bonfil et al., 2016). It has recently been used as an indicator, among others, to evaluate satellite based soil moisture retrievals as compared to ground-truth measurements (Kumar et al., 2018). These indicators may be useful for assessing the representative sample size as they reflect both mean and standard deviation, along with probability distributions without prior information (Cover and Thomas, 1991; Sim and Reid, 1999).

The purpose of this manuscript is to estimate the sample size required to accurately represent the modelled or estimated dataset. The purpose of this manuscript is not to estimate the sample size required to determine the 'true' value, or to determine the accuracy of the dataset as compared to a 'true' value.

It is proposed that validation of large remote sensing datasets, such as physical consistency and cross-comparison, can be analyzed through a representative sample size, which can be determined by the performance requirements of the dataset. This paper proposes that the CI and the maximum entropy probability of the sample dataset can define the threshold of the required sample dataset size to run these validation activities. A simple progressive sampling approach, used in machine learning and algorithm training, is adapted and used to determine the sample size to yield statistically significant results.

2. Materials and methods

The approach consists of four steps. First, the datasets are acquired (section 2.1). Second, the sampling schedule is defined and the samples are extracted (section 2.2). Third the performance measures defined and calculated for each sample (section 2.3). Last, the sample size at which convergence is achieved is detected (section 2.4).

2.1. The dataset

The approach was applied to six remote sensing-based datasets that cover continental Middle East and Africa (Fig. 1). Remote sensing-based products include actual evapotranspiration and interception (ETIa), net primary productivity (NPP), solar radiation (SR), reference evapotranspiration (RET), relative soil moisture index (SM) and normalized difference vegetation index (NDVI). The data used covers two spatial resolutions and two temporal resolutions. The resolution, image date, pixel count, image CI and sensor or data product used as input, for each image used is shown in Table 1.

All data was sourced from the Level 1 continental products in the WaPOR database version 1 (FAO, 2017). The ETIa, NPP and RET are sourced directly from the WaPOR portal (https://wapor.apps.fao.org/home/WAPOR_2/1), and the NDVI, SR and SM were provided by the WaPOR dataset producers. The WaPOR datasets are produced by the FRAME Consortium, led by eLEAF and comprised of The Flemish institute for technological research (VITO), International Institute for Geo-Information Science and Earth Observation at the University of Twente and WaterWatch. WaPOR undertakes gap filling, therefore all products are void of data gaps.

The ETIa, SM, NPP and NDVI are derived from the Moderate Resolution Imaging Spectroradiometer (MODIS)/Terra Surface Reflectance Daily L2G Global 250 m SIN Grid (MOD09GQ). The SR product uses the Digital Elevation Model (DEM) from the Shuttle Radar Topography Mission (SRTM) and transmissivity from the Meteosat Second Generation (MSG).

WaPOR ETIa, SM and NPP further relies on input from weather data (i.e. air temperature, relative humidity wind speed) which is obtained from Modern-Era Retrospective analysis for Research and Applications (MERRA). The weather data is resampled using a bilinear interpolation method to the 250 m resolution. The temperature is also resampled based on elevation data (FAO, 2018). The RET product is based only on the weather data and solar radiation.

Further, the CI performance criterion was applied to the ETIa product, and tested for different climate classes using a Köppen-Geiger

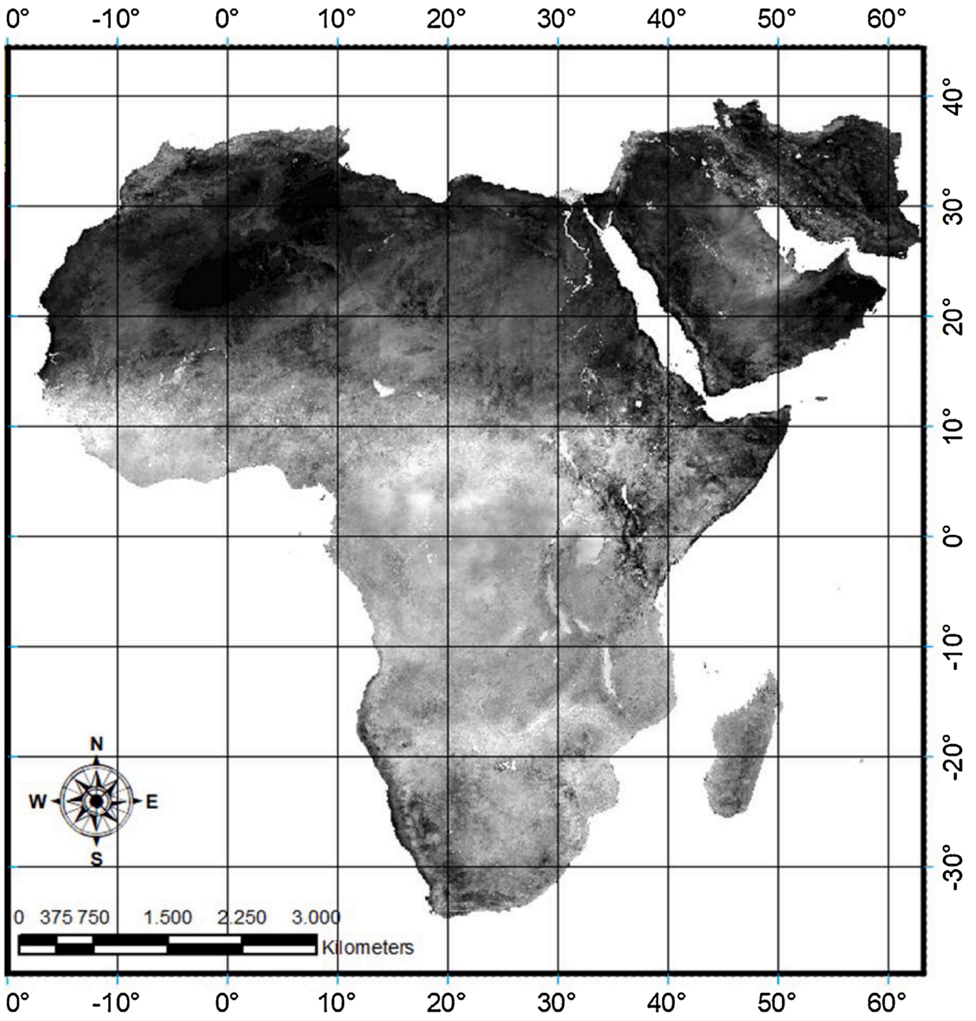


Fig. 1. The extent of the images used in the progressive sampling.

Table 1
Dataset properties for images used.

Data	Resolution	Image date	Pixel count	CI*	Data product/s	Sensor/s
ETIa	250 m|10-day	10 Apr 2009	1.22E + 09	1.78	MOD09GQ	MODIS; MERRA; MSG
SM	250 m|10-day	10 Apr 2009	1.22E + 09	0.84	MOD09GQ	MODIS
NPP	250 m|10-day	10 Apr 2009	1.22E + 09	2.81	MOD09GQ	MODIS; MERRA; MSG
NDVI	250 m|10-day	10 Apr 2009	1.22E + 09	1.27	MOD09GQ	MODIS
SR	250 m|1-day	18 Nov 2009	1.22E + 09	0.66	STRM (DEM)	MSG
RET	25 km|1-day	15 Dec 2009	2.38E+05	0.92	STRM (DEM)	MERRA; MSG

* CI is confidence interval as defined in Equation 2.

classification (Kottek et al., 2006). It is suggested that this approach can be used in evaluating the dataset for physical consistency, convergence should be achieved, or quantified, for each class or characteristic in which the physical consistency test is applied. For example, division of regions or classes can be based on hemisphere, climate or land use. The four major climate classes were: Arid (2.6E09 km^2), Equatorial (1.1E09 km^2), warm temperate (0.3E09 bil km^2) and humid continental (360, 450 km^2). The class arid, equatorial, warm temperate and humid continental are represented with all classes starting with B, A, C and D respectively.

2.2. *Sampling schedule*

The sample schedule in this research followed a geometric approach, due to the large size of the data (Estrada and Morales, 2004), using one geometric constant – a = 10, and two starting sample sizes, n0 = 100 and n0 = 300:

$$S_i = \left[a{\cdot}n0,\ \ a^2{\cdot}n0,\ a^3{\cdot}n0,\ \ldots \right) \text{ where } n0 = 100 \text{ and } n0 = 300$$

where Si is the sampling schedule, n0 is the starting sample size, a is the geometric factor and Ni is the sample size. Each sample increment, or sample size, is referred to as Ni. Each sample size was extracted 10 times (a sample set), with samples within a sample size referred to as Nj, Nj+1, Nj+2… Nj+9. This allows repetition of the test. The constant, a, was selected so the sample size can be identified quickly, considering the large size of the dataset (Table 1). The starting sample sizes were selected randomly as a way to reduce the aggressiveness of the approach. The sampling schedule tentatively reached Ni = 3000,000, which can be extended if convergence is not met. A random sampling method was used for all sampling schedules to ensures consistency. It was assumed that the results achieved for a single image can be extrapolated over time.

2.3. Sample extraction

All spatial samples were randomly generated in R software. One sample was generated for each test. Therefor there are 100 random samples per dataset (NPP, ETI etc), i.e. 10 randomly generated samples for each of the 10 sample sets. These random spatial point datasets were then used to extract the dataset values for each of the datasets. Where climate classification is considered, the climate class associated with each spatial point in the random spatial dataset was over-layed to extract the feature (climate class) of that point.

2.4. Performance measure

This study will use the CI and the differential entropy, or maximum entropy distribution, of the sample dataset (x) as indicators or performance criterion. Nominally, the acceptable performance for this case study was taken as 5% (ΔCIi,j = 5%). The acceptable entropy was defined as the entropy where the dependence on Nj is negligible. It is set nominally as ΔH(x) = 0.05.

The definition of the CI used as an indicator in this research is taken as (Clemmens and Burt, 1997):

$$CI_x = \pm 2CV_x$$

Where, CIx is the CI and CVx is the coefficient of variation. This CI definition is commonly applied in hydrology. It gives a measure of the CI relative to the magnitude of the expected value rather than the actual value which is found when using the z-coefficient (i.e. number of standard deviations) (Clemmens and Burt, 1997). The CVx is defined as:

$$CV_x = sd_x / \overline{m}_x$$

Where sd_x is the standard deviation of the sample dataset and $\overline{m}_x$ is the mean of the sample dataset. The sd_x is the standard statistical measure of variability. Although the CI is formally taken as the mean ±$2sd_x$, the CI defined in equation (4) provides a relative accuracy, with no units and often expressed as a percent.

The Principle of Maximum Entropy for continuous distributions, the differential entropy, is defined as (Jaynes, 2003, 1957):

$$H(x) = -\int_{-\infty}^{\infty} p(x)\ln p(x)dx$$

Where H(x) is the differential entropy and p(x) is the probability density function. This function applies to any probability density function that can be defined. The base of log is not important as long as it is uniform, as changing the base simply changes the scale of the entropy (Rajan et al., 2017). This requires a randomly generated sample dataset. The natural logarithm (ln) is used in this case. The entropy of the dataset will increase with an increasing population. The higher the entropy, the more information is given to that distribution. Therefore, when the marginal increase in entropy is negligible or minimal, little to negligible information can be gained by increasing the population size.

2.5. Detecting convergence

Detecting convergence requires statistical judgement on what performance is suitable. Once the required performance is determined, the suitability of the sample size can be determined. The Probably Close Enough Criteria (PCE) is a deduction procedure used in machine learning. It outputs an expression that has a high likelihood of closely approximating the expression to be learned (Valiant, 1984). Meaning that there is only a small chance that the mining algorithm could do better in training the algorithm using the entire database instead of the defined sample size. The PCE defines the suitable sample size as (John et al., 1996; Provost et al., 1999):

$$(acc(N_{i+1}) - acc(N_i) > \Delta E) \leq \delta$$

Where acc(N) refers to the accuracy (acc) of the sample size (Ni). ΔE refers to the acceptable increase in accuracy (or marginal increase) and δ is the probability that the maximum accuracy will be exceeded on any run, therefore, satisfying the accuracy requirement for each run of a sample size and any increase in sample size.

The PCE criterion is adapted to determine the suitable sample size with the selected performance measures. The marginal increase in performance, referred to as ΔE in equation 5, is determined by calculating the statistical variation in CI, or the differential entropy with increasing sample size steps (Ni):

$$\left(CI_{i+1,j} - CI_{i,j} \leq \Delta CI_i\right) \text{ and } \left(H(x)_{i+1,j} - H(x)_{i,j} \leq \Delta H(x)_i\right)$$

Where the ΔCIi and ΔH(x)i is the marginal change of CI and H(x) with increasing sample size. If the difference in the CI is greater than the acceptable CI (ΔCIi) for any sample size increase, Ni, the sample size is rejected, and the sample size is increased.

The probability of that the maximum performance will be exceeded on any run, referred to as δ in equation 5, is considered by running the test on each sample size several times. Equation (6) is adapted to refer to the statistical variation in CI, or the differential entropy, of any sample, Nj, within the sample size, Ni, becoming:

$$\left(CI_{i,j+1} - CI_{i,j} \leq \Delta CI_j\right) \text{ and } \left(H(x)_{i,j+1} - H(x)_{i,j} \leq \Delta H(x)_j\right)$$

Where the ΔCIj and ΔH(x)j is the range of performance of samples Nj for sample size Ni. If the difference in the CI is greater than the acceptable CI (ΔCIj) for any sample, Nj, the sample size is rejected and sample size is increased. Both equations 8 and 9 need to be met for the sample size to be considered suitable. This was undertaken for the entire sample data set.

The expected trend of the CI and the entropy for an infinitely large, positive, continuous dataset, with increasing sample size, is shown in Fig. 2. The black crosses represent the CI or differential entropy for each sample Nj for a given sample size Ni, within the sampling schedule, Si. The black lines show the expected maximum (plotted for CI and entropy) and minimum (plotted for CI only) values of the sample set. The decreasing range in performance metric values for increasing sample size reflects convergence. The CI is expected to converge at a mean value, while entropy is expected to converge to a maximum value.

2.6. Cross-comparison

An internal cross-compared was undertaken to determine if relationships between datasets remained with increasing sample size. There is a strong link between ETIa and NPP (Blatchford et al., 2019), which is therefor used to assess the relationship variation with increasing sample size. Correlation was selected as the metric. The correlation was estimated for all samples in all sample sets.

3. Results

The CI for the full set of sampling schedules and for all datasets, are shown in Fig. 3. The sample size is plotted on the x-axis, using a logarithmic scale, and the CI is plotted on the y-axis. The CI shows the most variation at low sample sizes and converges as the sample size increases. The range of CI is decreasing with increasing sample size for all samples to the CI value seen in Table 1. The rate of convergence is also greatest when the sample size is low, and decreases with increasing sample size. There are 3 occasions where the range in CI increases with an increasing step size: the ETIa CI range is greater when Ni = 100,000 than when Ni = 30,000, the RET CI range is greater at Ni = 1000 than when Ni = 300 and the NPP when Ni = 100.000 and Ni = 300.000. This type of variation before convergence is expected as extreme values are expected to have a greater influence in the data distribution and CI when the sample size is small.

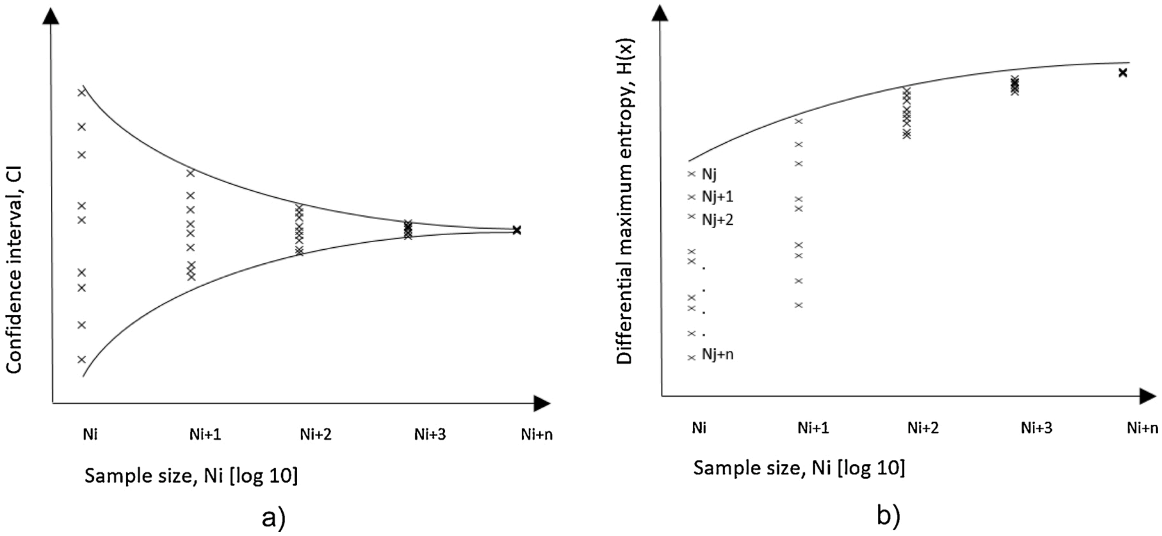


Fig. 2. Expected learning curve trend of the CI (a) and the entropy (b) for repeated tests and increasing sample size. Black lines indicate trends. The black line shows the expected trend of variation for increasing sample size.

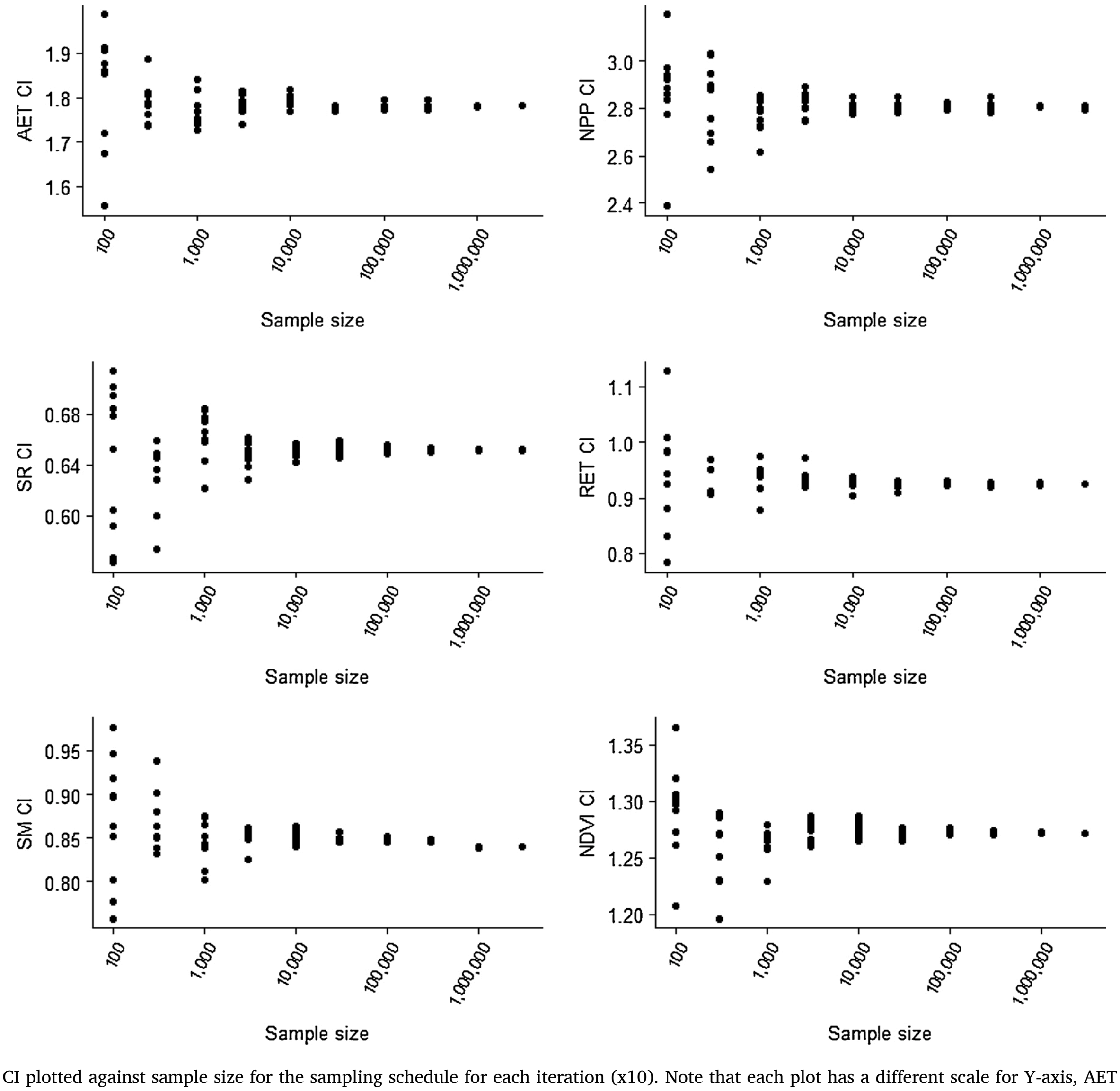


Fig. 3. CI plotted against sample size for the sampling schedule for each iteration (x10). Note that each plot has a different scale for Y-axis, AET is ETIa.

The smallest CI range (ΔCIj) occurs at Ni = 3000,000 for all datasets, which is less than 0.05 % for all datasets. The SR, SM and NDVI datasets reached a CI range of less than 5% at Ni = 3000, with CI ranges of 3.34 %, 3.57 % and 2.51 % respectively. The ETIa and RET datasets both reached a CI range of less than 5% at Ni = 10,000, with CI ranges of 4.63 % and 3.41 % respectively. The NPP reached a CI range of less than 5% at Ni = 100,000 or 3.16 % of total sample size.

The CI range with increasing step size (ΔCIi) is less than 5% at a sample size of 30,000 for NPP. The ΔCIi condition is met at the same sample size that ΔCIj is met for ETIa, SR, SM, NDVI and RET. Therefore for ΔCIi,j range of 5% is met at Ni = 3000 for SR and SM and at Ni = 10,000 for ETIa, RET and NDVI. While for NPP it is met at Ni = 100,000.

The distribution of the ETIa samples sets Ni = 300, Ni = 3000 and Ni = 30,000, and the differential entropy plotted against the sample size is shown in Fig. 4. The 1—day ETIa is plotted on the x-axis and the density for a given ETIa is plotted on the y-axis. Each line represents a single sample Nj, for a given sample size, Ni. The variation in the distribution functions is greatest when the sample size is low, Ni = 300. The range in the amplitude of both peaks varies, and with high ETIa values. The greatest variation for Ni = 3000 is seen at the peaks and troughs. The variation for Ni = 30,000 is only visible at the second peak, but the variation is less visible than the smaller sample sizes. The higher sample sizes have a greater convergence. This is represented by the merging distribution functions. While different datasets had a different distribution function, all showed converging distribution functions with increasing sample size.

Fig. 5 shows the entropy of all samples for all datasets. The entropy values converge with increasing sample size for all datasets. All datasets show increasing entropy with increasing sample size. Entropy values are sometime higher at lower sample sizes as compared to higher sample sizes. However, the values are converging to higher mean values for each increasing sample size. The range of CI values are highest for all datasets at Ni = 100 or Ni = 300. The rate of convergence is greatest when the sample size is low, and decreases with increasing sample size.

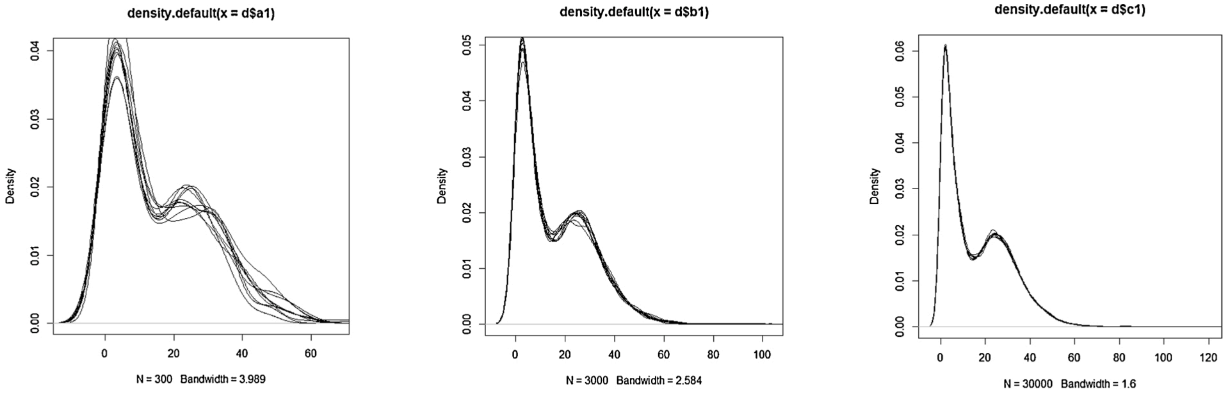
There are only 2 occasions where the range in CI increases with an increasing step size; the ETIa entropy range is greater when Ni = 3000 than when Ni = 1000 and the NDVI entropy range is greater at Ni = 1000 than when Ni = 300. The minimum entropy range ΔH(x)j occurs at Ni = 3000,000 and is less than 0.003, for all datasets. This is followed by Ni = 1,000,000, where the range is entropy is less than 0.006 for all datasets. At Ni = 300,000, Ni = 100,000, Ni = 30,000 and Ni = 10,000 the ΔH(x)j is less than 0.04, 0.05, 0.06 and 0.08 respectively. For sample sizes less than Ni = 3000, the entropy ranges are much larger, ranging from RET ΔH(x)j = 0.05 when Ni = 3000, to NPP ΔH(x)j = 0.68 when Ni = 100. The ΔH(x)i ≤0.05 performance indicator is met at Ni = 3000 for RET, Ni = 10,000 for SR and SM and Ni = 30,000 for NDVI, NPP and ETIa. The ΔH(x)j ≤0.05 performance indicator is met at Ni = 3000 for RET, Ni = 10,000 for SM, SR and RET, at Ni = 30,000 for NPP and ETIa. Therefore for ΔH(x)i,j condition is met at Ni = 3000 for RET, Ni = 10,000 for SR and SM, at Ni = 30,000 for NPP, ETIa and NDVI.

The CI and entropy precision increases for both the CI and the differential entropy with increasing sample size. This is reflected in the increasing density of the cluster for increasing sample sizes. The CI values are converging to a single, central value. Comparatively, the differential entropy values are converging to a higher value. Differential entropy increasing coupled with decreasing variation for increasing sample size conforms to expectations (Fig. 1).

The ETIa dataset was used as example to show the performance of the CI indicator when using for climate classes. The CI range for each climate class and each sample size is shown in Fig. 6. The classes with the largest CI interval ranges have the deepest saturation of red and the classes with the lowest ranges have the deepest saturation of green. The classes where the sample size was not large enough for any of the samples in the sample set to provide information on the CI are grey. The smallest sample size, Ni = 100, shows the greatest CI variation for all classes, with the CI range frequently exceeding 2 (or 200 %). This includes classes with the largest representation, such as the arid desert hot class (BWh). As the sample size increases the CI range decreases for most classes. Exceptions occur for some sample size increments where the class has a faction of total area of less than 1%. When the sample size reached Ni = 300,000, the CI is below the previous sample size, Ni = 100,000, for all classes. At Ni = 100,000, one class has greater CI than the class before (greater than CIj >0.05). This is for the class with the smallest area, temperate dry warm summer (Csb), and the number of sample points representing this class is still <100. At Ni = 3000, all classes had samples to estimate the CI for the entire sample set.

When the sample size is Ni = 3000 the CI range is ΔCIj <2 for all classes and at 1000, the CI range is ΔCIj >2 on one occasion, cold (continental) dry warm summer (Dsb). When the sample size is Ni = 30,000, one class has a ΔCIj = 1 and all other classes have a range in CI of less than 0.8. Four classes have a range in CI with ΔCIj <0.05 and seven classes with ΔCIj <0.15. When the sample size is Ni = 100,000 the maximum CI range is ΔCI = 0.51. Four classes have a range in CI with ΔCI <0.05 and 12 have a ΔCI<0.15. When the sample size is Ni = 300,000 the maximum CI range is ΔCI = 0.34. Eight classes have a range in CI with ΔCIj<0.05 and 14 (of 16) have a ΔCIj<0.15. The samples that exceed ΔCI<0.15 are from the humid continental classes (cold (continental) dry hot summer (Dsa) and Dsb).

When applying the test to only the major classes - arid, equatorial temperate and humid continental – the same overall CI trend is observed, a decreasing ΔCIj for all major classes with increasing sample size. When the sample size is Ni = 3000, one class has a range in CI of



(a) Ni=300 (b) Ni=3,000 (c) Ni=30,000

Fig. 4. Density functions of ETIa sample size sets, a) Ni = 300, b) Ni = 3000 and c) Ni = 30,000.

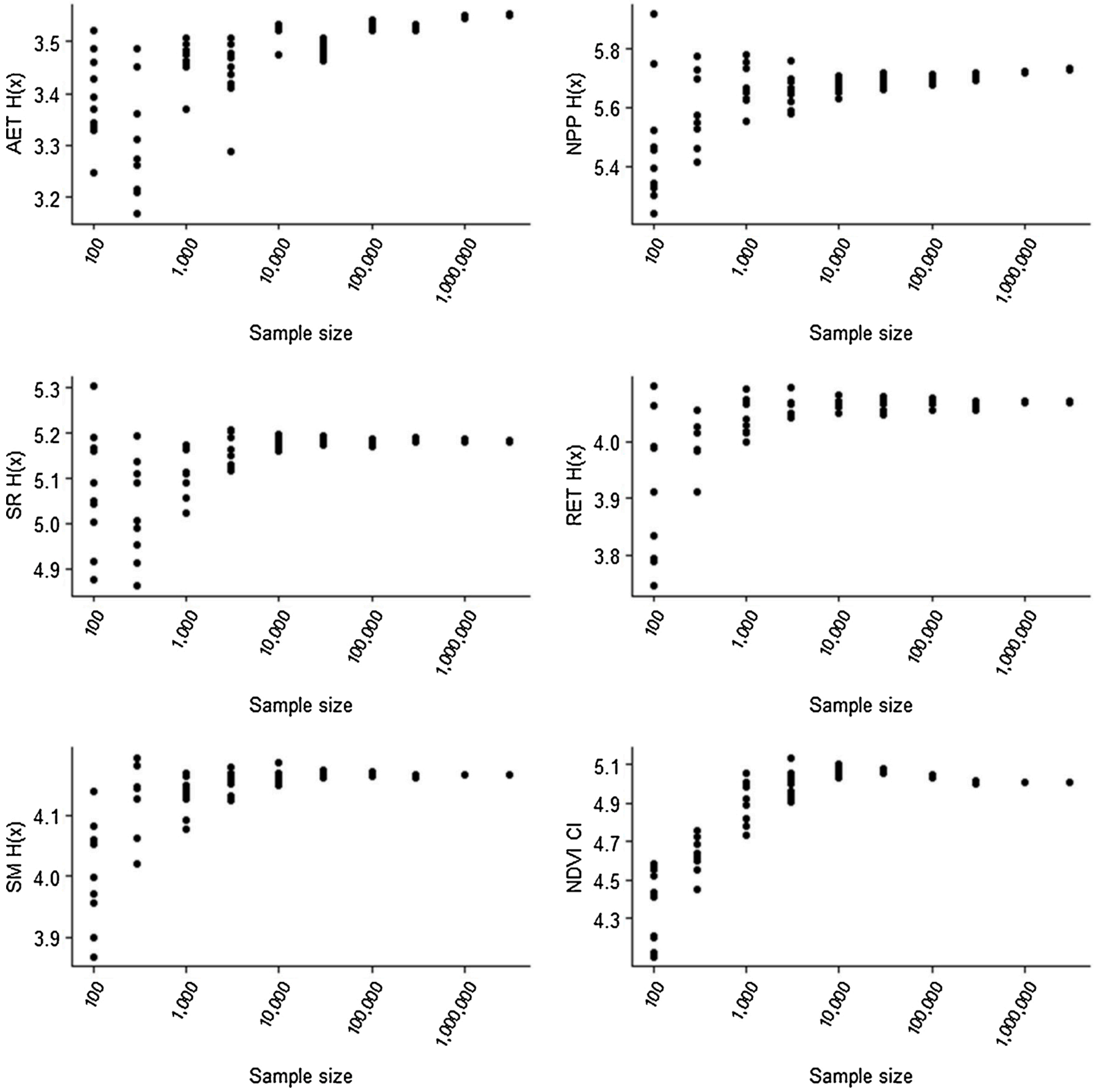


Fig. 5. Entropy, H(x), plotted against sample size for the sampling schedule for each iteration (x10). Note that each plot has a different scale for Y-axis; AET is ETIa.

$\Delta \mathrm{CIj} < 0.05$ (equatorial) and two others have a $\Delta \mathrm{CIj} < 0.15$ (equatorial and arid). When the sample size is $\mathrm{Ni} = 300{,}000$ the maximum CI range is $\Delta \mathrm{CIj} = 0.34$. When the sample size is $\mathrm{Ni} = 10{,}000$ up to and including $\mathrm{Ni} = 100{,}000$, one class has a range in CI of $\Delta \mathrm{CI} < 0.05$ (equatorial) and three have a $\Delta \mathrm{CIj} < 0.15$ (equatorial, arid and warm temperate). When the sample size reaches $\mathrm{Ni} = 300{,}000$, three classes have a range in CI of $\Delta \mathrm{CIj} < 0.05$ (equatorial, arid and warm temperate). The minimum $\Delta \mathrm{CIj}$ in humid continental class is $\Delta \mathrm{CIj} = 0.19$ and occurs at $\mathrm{Ni} = 300{,}000$. The humid continental has the smallest area, <1% of the total area, and therefore has the smallest number of representative sample points.

Fig. 7 shows the correlation between the AET and NPP with increasing sample size. Similar to the CI for each dataset, the correlation shows the most variation at low sample sizes and converges as the sample size increases. This shows that, like CI and H(x), negligible further insight into the relationship is gained beyond a certain sample size. In this case depending on the users preferred margin of error, this is likely to fall between a sample size of 30,000, where correlation ranges from 0.41 to .45, and 300,000, where correlation ranges from 0.41-0.43.

4. Discussion

In this research, a new and operative methodology was proposed to define a representative sample for arbitrarily large, continuous datasets, using a progressive sampling approach combined with two performance indicators. The purpose being to increase efficiency of validation and quality assessment tasks where the entire dataset has previously been used. The results showed that the assessed datasets in the continent of Africa and the Middle East, without classification of zones, can suitably be represented by a small fraction dataset as the performance condition of both indicators, CI and entropy, was met at $\mathrm{Ni} = 10{,}000$ for RET, SM and SR, $\mathrm{Ni} = 30{,}000$ for ETIa and NDVI and at $\mathrm{Ni} = 100{,}000$ for NPP. This represents 0.01 % for the of the total dataset size of the 250 m resolution datasets, and 0.41 % of the total dataset size 25 km resolution dataset (Table 1).

Though no directly comparable study exists, several studies in other fields that use progressive sampling for to increase training efficiency of discrete datasets exist. Six studies that look at a combined 22 different datasets, including land cover type (Lazarevic and Obradovic, 2001; Peng et al., 2004), traffic data (Umarani and Punithavalli, 2011), waveform (Lazarevic and Obradovic, 2001; Ng and Dash, 2006; Peng

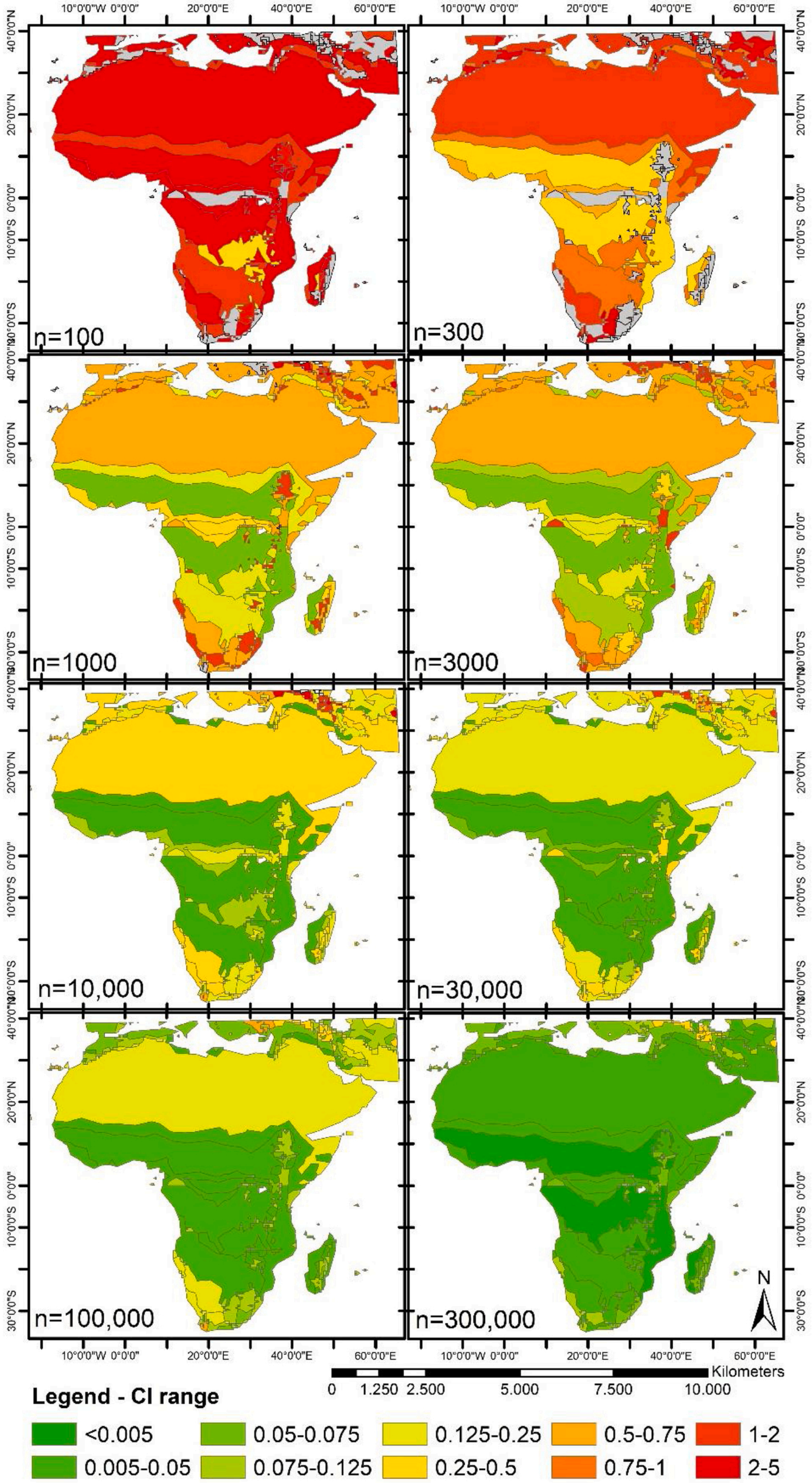


Fig. 6. The CI range for each sample set for ETIa using Köppen-Geiger climate classes for all sample sizes (Ni).

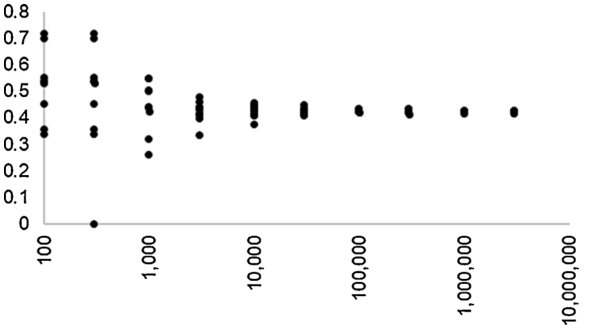


Fig. 7. Correlation between AET and NPP plotted against sample size for the sampling schedule for each iteration (x10).

et al., 2004), simulated data (ElRafey and Wojtusiak, 2017; Umarani and Punithavalli, 2011), wine quality data (ElRafey and Wojtusiak, 2017), with varying number of categories or attributes. The effective sample size was determined by each author and is not related to the indicators selected in this study. Irrelevant of the parameter, the sampling schedule or the number of categories, we found a power relationship between the effective sample size per category among these studies (Fig. 8). Although a small pool of data, the power relationship between the total dataset size and the sample size required for effective algorithm training has a good coefficient of determination ($R^2 = 0.76$). If this power relation is applied to the dataset size used in this research, as if there is one category, an effective sample size is extrapolated to be 0.2 % and 3% of total dataset size for the 250 m and 25 km resolution datasets respectively, which is similar to our results. However, none of these datasets are arbitrarily large, or continuous and in theory this should the sample size would stop increasing beyond a certain point (Cherkassky et al., 1999). The entropy performance measure accounts for this, as the negligible marginal gain in new information with increasing sample size is reflected in a negligible marginal gain in entropy.

While the proposed approach should remain effective regardless of different state variable with different spatial-temporal heterogeneity, the determined sample size may vary. This was seen between the six observed datasets. The required sample size should be particularly influenced by the dataset complexity rather than specifically dataset resolution (unless the resolution is increasing dataset complexity). For example, in non-discrete applications of progressive sampling, samples with increasing number of attributes is associated with increasing sample size, as the effective sample size was frequently determined per category.

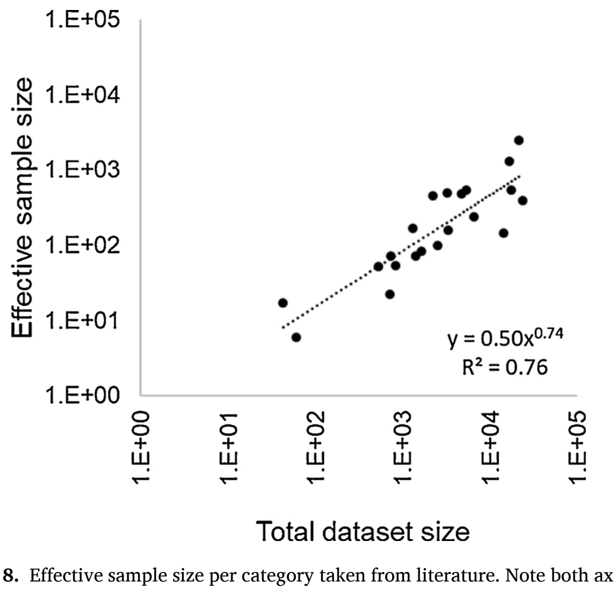


Fig. 8. Effective sample size per category taken from literature. Note both axis' use log scale.

While the CI indicator can be easily determined based on user performance requirements, differential maximum entropy is less obvious. This study used a nominal requirement of $\Delta H(x) < 0.05$. If the acceptable value decreased to $\Delta H(x) < 0.02$ the required sample size required increases significantly. For example, the ETIa sample size requirement becomes Ni = 300,000 as compared to Ni = 30,000. Alternatively, if it is increased to $\Delta H(x) < 0.10$ the acceptable sample size is Ni = 1000 for this indicator (although it would still need to meet the CI indicator requirement). This not only highlights the importance of selecting an acceptable entropy increment suitable to user application, but also highlights the sensitivity of the entropy indicator.

The sample size required to reach the acceptable performance increased once climate classification was introduced. When the sample size reached 300,000 only 8 reached the acceptable ΔCI. A sample size of 30,000, achieved the acceptable performance for only 4 and 7 classes respectively. This suggests that the sample size included in the sampling schedule does not meet the PCE criterion when considering all classes. Similarly, the PCE criterion was not met when only the major climate classes were considered. The smallest classes, humid continental classes, did not meet the acceptable error for the entire sampling schedule. This was a result of the small representation of humid continental class. Rather than increase the sample size, it may be more appropriate to use methods such as a stratified random sample, or a progressive boosting to optimize sample size and account for imbalanced data (Lazarevic and Obradovic, 2001; Soleymani et al., 2018). This would align with approaches used in land cover classification where a minimum sample size per class is often defined (EFTAS, FAO 2015). This highlights the importance of selecting a sampling approach that suits the user needs. For studies focused on the Middle East and Africa, a lower confidence in the humid continental class, which is located predominately in Europe, may be acceptable.

While the approach has only been tested on the region of Africa and the Middle East to climate data, it is possible to extend the approach to other regions or for other categories. While the convergence point for both indicators is expected to change based on data distribution and on how the data is categorized (e.g, climate of land cover type), the progressive sampling approach should still be valid, and is expected though the sampling size may increase based on data complexity resulting from increased environment, land cover climate, topography types etc. In the case of added complexity or categories a complexity measure, as applied with entropy to soil moisture retrievals (Kumar et al., 2018), may also by useful.

This study used spatially continuous datasets (no data gaps). In cases where data gaps exist the sampling schedule should be applied to pixels where data exists (exclude no data), or the sample size will increase, however, dependent on dataset, this will result in areas with large data gaps missing from the evaluation.

The two indicators selected cover both the mean and standard deviations, through the CI of the dataset, and the distribution of the dataset through the differential entropy. They are useful indicators in the absence of true values, which is relevant for large spatial datasets with little ground-truth information available. These indicators are therefore not intended to define the accuracy of the dataset itself, but how accurately the sample represents the dataset. The condition of both indicators should be met when defining the suitable sample size. This was seen in this example when the CI acceptable criterion was met before the differential entropy acceptable was met. The usefulness of the approach is that the performance criterion (i.e. CI and entropy) can be defined dependent on the application and therefore accuracy requirement of the user.

The selected indicators selected are not suitable for all continuous datasets, as not all data behaves as a continuous dataset. For example, although precipitation is a continuous variable, the number of zero rain points will cause the dataset to behave as a discrete-continuous distribution (Friederichs and Hense, 2007). For datasets that behave discretely, it may be useful to use the two-step approach that is commonly used in precipitation validation, whereby the product is first validated categorically and then quantitatively (Wilks, 2006). Though, this shows limitations as for precipitation this will no longer represent dry or arid regions well.

Finally, a similar approach could, in theory, be applied to determining sample size for comparison against ground-truth measurements. Many limitations are imposed on ground-truth data collection, such as sample design (due to access constraints) and resource limitations. However, it could guide a best practice or an or objective for a community, for example fluxnet, where data collection is less intensive and can be collected without experts, for example citizen science.

5. Conclusions

The progressive sampling approach combined with CI and differential maximum entropy performance measures, can be applied to determine a suitable sample size for physical consistency and cross-comparison tests of a continuous, arbitrarily large remote sensing-based datasets. The approach showed that the amount of data required to represent large datasets (1.22E + 09 pixels) datasets is comparatively small (10,000–100,000 pixels) and therefore can significantly reduce computing time and resources. This can be used to run initial tests or product analysis. It is suggested that using a representative sample, rather than the whole dataset, can effectively balance insight to the quality of the dataset and reduce processing efforts required in validation procedures, such as continental cross-comparisons, that are computationally exhaustive. This will become even more useful as dataset resolution increases.

Funding

This work was partly funded by the United Nations Food and Agricultural Organization (FAO) as part of the Water Productivity Open Access Portal (WaPOR) project, in which we undertook the validation efforts. The WaPOR validation efforts were undertaken under contract number 2016/HQR/NRL-CPA320260 (FRAME project).

CRediT authorship contribution statement

Megan L. Blatchford: Conceptualization, Formal analysis, Methodology, Visualization, Writing - original draft, Software. **Chris M. Mannaerts:** Conceptualization, Formal analysis, Supervision, Writing - review & editing. **Yijian Zeng:** 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.

Acknowledgments

We would like to acknowledge the FRAME consortium members, eleaf and VITO, for providing the dataset and the ongoing support of the validation activities for the WaPOR project.

Appendix A. Supplementary data

Supplementary material related to this article can be found, in the online version, at doi:https://doi.org/10.1016/j.jag.2020.102235.

References

Blatchford, M.B., Mannaerts, C.M., Zeng, Y., Nouri, H., Karimi, P., 2019. Status of accuracy in remotely sensed and in-situ agricultural water productivity estimates: a review. Remote Sens. Environ. 234, 111413 https://doi.org/10.1016/j.rse.2019.111413.

Burt, C.M., Clemmens, A.J., Strelkoff, T.S., Solomon, K.H., Bliesner, R.D., Hardy, L.A., Howell, T.A., Eisenhauer, D.E., 1997. Irrigation performance measures: efficiency and uniformity. J. Irrig. Drain. Eng. 123, 423–442. https://doi.org/10.1061/(ASCE)0733-9437(1997)123:6(423).

Chen, C., Li, Y., 2013. A Robust Multiquadric Method for Digital Elevation Model Construction. Math. Geosci. 45, 297–319. https://doi.org/10.1007/s11004-013-9451-8.

Cherkassky, V., Xuhui, S., Mulier, F.M., Vapnik, V.N., 1999. Model complexity control for regression using VC generalization bounds. IEEE Transactions on Neural Networks. IEEE, pp. 1075–1089. https://doi.org/10.1109/72.788648.

Clemmens, A.J., Burt, C.M., 1997. Accuracy of irrigation efficiency estimates. J. Irrig. Drain. Eng. 123, 443–453. https://doi.org/10.1061/(ASCE)0733-9437(1997)123:6(443).

Congalton, R.G., Green, K., 2009. Assessing the accuracy of remotely sensed data - principles and practices. Int. J. Appl. Earth Obs. Geoinf. 11, 183. https://doi.org/10.1016/j.jag.2009.07.002.

Cover, T.M., Thomas, J.A., 1991. Elements of information theory. Elem. Inf. Theory 1–748. https://doi.org/10.3390/e7040253.

EFTAS, FAO, 2015. Protocol for Land Cover Validation. Rome, Italy.

ElRafey, A., Wojtusiak, J., 2017. Recent advances in scaling-down sampling methods in machine learning. Wiley Interdiscip. Rev. Comput. Stat. 9 https://doi.org/10.1002/wics.1414.

Estrada, A., Morales, E.F., 2004. NSC: a New progressive sampling algorithm. Proceedings of the Workshop: Machine Learning Learning for Scientific Data Analysis (Iberamia) 335–344.

FAO, 2017. Using Remote Sensing in Support of Solutions to Reduce Agricultural Water Productivity Gaps. Database Methodology: Level 1 Data. WaPOR Beta Release.

FAO, 2018. WaPOR database methodology: level 1. Remote Sensing for Water Productivity Technical Report: Methodology Series. FAO, Licence: CC BY-NC-SA 3.0 IGO, Rome, Italy.

Figueroa, R.L., Zeng-Treitler, Q., Kandula, S., Ngo, L.H., 2012. Predicting sample size required for classification performance. BMC Med. Inform. Decis. Mak. 12, 8. https://doi.org/10.1186/1472-6947-12-8.

Foody, G.M., 2002. Status of land cover classification accuracy assessment. Remote Sens. Environ. 80, 185–201. https://doi.org/10.1016/S0034-4257(01)00295-4.

Friederichs, P., Hense, A., 2007. Statistical Downscaling of Extreme Precipitation Events Using Censored Quantile Regression. Mon. Weather Rev. June 2007, 2365–2378. https://doi.org/10.1175/MWR3403.1.

Fukunaga, K., Hayes, R.R., 1989. Effects of sample size in classifier design. IEEE Trans. Pattern Anal. Mach. Intell. 11, 873–885. https://doi.org/10.1109/34.31448.

Gu, B., Liu, B., Hu, F., Liu, H., 2001. Efficiently determining the starting sample size for progressive sampling. Mach. Learn. ECML 2001, 192–202. https://doi.org/10.1007/3-540-44795-4_17.

Gupta, K.K., Attri, J.P., Singh, A., Kaur, H., Kaur, G., 2016. Basic concepts for sample size calculation: critical step for any clinical trials! Saudi J. Anaesth. 10, 328–331. https://doi.org/10.4103/1658-354X.174918.

Heydari, S.S., Mountrakis, G., 2018. Effect of classifier selection, reference sample size, reference class distribution and scene heterogeneity in per-pixel classification accuracy using 26 Landsat sites. Remote Sens. Environ. 204, 648–658. https://doi.org/10.1016/j.rse.2017.09.035.

Jaynes, E.T., 1957. Information Theory and Statistical Mechanis. Phys. Rev. 106, 620–630.

Jaynes, E.T., 2003. Probability Theory: The Logic of Science. Cambridge University Press, Cambridge, UK.

John, G., Langley, P., John, H., 1996. Static Versus Dynamic Sampling for Data Mining. Kdd 367–370. https://doi.org/10.1007/s00221-011-2539-9.

Kim, S.Y., 2009. Effects of sample size on robustness and prediction accuracy of a prognostic gene signature. BMC Bioinformatics 10, 4–7. https://doi.org/10.1186/1471-2105-10-147.

Kirby, A., Gebski, V., Keech, A.C., 2002. Determining the sample size in a clinical trial. Med. J. Aust. 177, 256–257. https://doi.org/10.5694/j.1326-5377.2003.tb05241.x.

Kottek, M., Grieser, J., Beck, C., Rudolf, B., Rubel, F., 2006. World map of the Köppen-Geiger climate classification updated. Meteorol. Zeitschrift 15, 259–263. https://doi.org/10.1127/0941-2948/2006/0130.

Kumar, S.V., Dirmeyer, P.A., Peters-Lidard, C.D., Bindlish, R., Bolten, J., 2018. Information theoretic evaluation of satellite soil moisture retrievals. Remote Sens. Environ. 204, 392–400. https://doi.org/10.1016/j.rse.2017.10.016.

Lachin, J.M., 1981. Introduction to sample size determination and power analysis for clinical trials. Control. Clin. Trials 2, 93–113. https://doi.org/10.1016/0197-2456(81)90001-5.

Last, M., 2009. Improving data mining utility with projective sampling. Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min 487–495. https://doi.org/10.1145/1557019.1557076.

Lazarevic, A.O., Obradovic, Z., 2001. Data reduction using multiple models integration. In: de Raedt, L., Siebes, A. (Eds.), Principles of Data Mining and Knowledge Discovery: 5th European Conference. PKDD 2001, Springer, Berlin, pp. 301–313.

Loew, A., Bell, W., Brocca, L., Bulgin, C.E., Burdanowitz, J., Calbet, X., Donner, R.V., Ghent, D., Gruber, A., Kaminski, T., Kinzel, J., Klepp, C., Lambert, J., Schaepman-Strub, G., Schröder, M., 2017. Validation practices for satellite based earth observation data across communities. Rev. Geophys. https://doi.org/10.1002/2017RG000562.

Luo, G., 2016. A review of automatic selection methods for machine learning algorithms and hyper-parameter values. Netw. Model. Anal. Heal. Informatics Bioinforma. 5, 1–16. https://doi.org/10.1007/s13721-016-0125-6.

Makarovic, B., 1973. Progressive sampling for digital terrain models. ITC J. 3, 145–153.

Meek, C., Thiesson, B., Heckerman, D., 2002. The learning-curve sampling method applied to model-based clustering. J. Mach. Learn. Res. 2, 397–418. https://doi.org/10.1162/153244302760200678.

Morisette, J.T., Privette, J.L., Justice, C.O., 2002. A framework for the validation of MODIS Land products. Remote Sens. Environ. 83, 77–96. https://doi.org/10.1016/S0034-4257(02)00088-3.

Ng, W., Dash, M., 2006. An evaluation of progressive sampling for imbalanced data sets. In: Proc. - IEEE Int. Conf. Data Mining. ICDM, pp. 657–661. https://doi.org/10.1109/ICDMW.2006.28.

Payton, M.E., 1996. Confidence intervals for the coefficient of variation. Conf. Appl. Stat. Agric. https://doi.org/10.4148/2475-7772.1320.

Peng, K., Obradovic, Z., Vucetic, S., 2004. Towards efficient learning of neural network ensembles from arbitrarily large datasets. Front. Artif. Intell. Appl. 110, 623–627.

Provost, F., Jensen, D., Oates, T., 1999. Efficient progressive sampling. In: Proc. Fifth ACM SIGKDD Int. Conf. Knowl. Discov. Data Min. KDD' 99, pp. 23–32. https://doi.org/10.1145/312129.312188.

Rajan, A., Kuang, Y.C., Ooi, M.P.-L., Demidenko, S.N., 2017. Moments and maximum entropy method for expanded uncertainty estimation in measurements. IEEE 3–8.

Rauhala, U.A., 1989. Compiler positioning system: an array algebra formulation of digital photogrammetry. Photogramm. Eng. Remote Sens. 55, 317–326.

Reed, G.F., Lynn, F., Meade, B.D., 2002. Quantitative Assays 9, 1235–1239. https://doi.org/10.1128/CDLI.9.6.1235.

Santamaría-Bonfil, G., Fernández, N., Gershenson, C., 2016. Measuring the complexity of continuous distributions. Entropy 18. https://doi.org/10.3390/e18030072.

Sarkar, A., Guo, J., Siegmund, N., Apel, S., Czarnecki, K., 2016. Cost-efficient sampling for performance prediction of configurable systems. In: Proc. - 2015 30th IEEE/ACM Int. Conf. Autom. Softw. Eng. ASE 2015, pp. 342–352. https://doi.org/10.1109/ASE.2015.45.

Schectman, O., 2013. Methods of clinical epidemiology. Methods Clin. Epidemiol. 33–49. https://doi.org/10.1007/978-3-642-37131-8.

Schneider, P., Ghent, D., Prata, F., Corlett, G.K., Remedios, J.J., 2012. AATSR Validation: LST Validation Protocol. ESA Contract Number: 19054/05/NL/FF. Report for the European Space Agency.

Shannon, C.E.C., 1948. A mathematical theory of communication. Bell Syst. Tech. J. 27, 379–423. https://doi.org/10.1145/584091.584093.

Sim, J., Reid, N., 1999. Statistical inference by confidence intervals: issues of interpretation and utilization. Phys. Ther. 79, 186–195. https://doi.org/10.1093/ptj/79.2.186.

Smets, B., Lacaze, R., Freitas, S.C., Jann, A., Calvet, J.C., Camacho, F., Baret, F., Paulik, C., d'Andrimont, R., Tansey, K., 2013. Operating The Copernicus Global Land Service. others ESA Spec. Publ. 722, 66.

Soleymani, R., Granger, E., Fumera, G., 2018. Progressive Boosting for Class Imbalance and Its Application to Face Re-Identificatio. Expert Syst. Appl. 101, 271–291. https://doi.org/10.1016/j.eswa.2018.01.023.

Steduto, P., Todorovic, M., 2001. The Agro-Ecological Charac- terisation of Apulia Region. Methodology and Experience 34, 143–164.

Stein, A., Ettema, C., 2003. An overview of spatial sampling procedures and experimental design of spatial studies for ecosystem comparisons. Agric. Ecosyst. Environ. 94, 31–47. https://doi.org/10.1016/S0167-8809(02)00013-0.

Umarani, V., Punithavalli, M., 2011. Analysis of the progressive sampling-based approach using real life datasets. Open Comput. Sci. 1, 221–242. https://doi.org/10.2478/s13537-011-0016-y.

Valiant, L.G., 1984. A theory of the learnable. Commun. ACM 27, 1134–1142. https://doi.org/10.1145/1968.1972.

Warmuth, M.K., Liao, J., Ratsch, G., Mathieson, M., Putta, S., Lemmen, C., 2003. Active learning with support vector machines in the drug discovery process. J. Chem. Inf. Comput. Sci. 43, 667–673.

Wilks, D., 2006. Forecast verification. In: Wilks, D. (Ed.), Statistical Methods in the Atmospheric Sciences. Elsevier, pp. 255–335.

Young, K.D., Lewis, R.J., 1997. What is confidence? Part 1: the use and interpretation of confidence intervals. Ann. Emerg. Med. 30, 307–310. https://doi.org/10.1016/S0196-0644(97)70166-5.

Zeng, X., Luo, G., 2017. Progressive sampling-based Bayesian optimization for efficient and automatic machine learning model selection. Heal. Inf. Sci. Syst. 5, 2. https://doi.org/10.1007/s13755-017-0023-z.

Zeng, Y., Su, Z., Calvet, J.C., Manninen, T., Swinnen, E., Schulz, J., Roebeling, R., Poli, P., Tan, D., Riihelä, A., Tanis, C.M., Arslan, A.N., Obregon, A., Kaiser-Weiss, A., John, V. O., Timmermans, W., Timmermans, J., Kaspar, F., Gregow, H., Barbu, A.L., Fairbairn, D., Gelati, E., Meurey, C., 2015. Analysis of current validation practices in Europe for space-based climate data records of essential climate variables. Int. J. Appl. Earth Obs. Geoinf. 42, 150–161. https://doi.org/10.1016/j.jag.2015.06.006.

Zeng, Y., Su, Z., Barmpadimos, I., Perrels, A., Poli, P., Boersma, K.F., Frey, A., Ma, X., de Bruin, K., Goosen, H., John, V.O., Roebeling, R., Schulz, J., Timmermans, W., 2019. Towards a traceable climate service: assessment of quality and usability of essential climate variables. Remote Sens. 11, 1186. https://doi.org/10.3390/rs11101186.