

Random Point and Area Sampling in Geospatial Accuracy Validation: A Literature Review

Introduction

Validation of image classification results (e.g. land-cover maps derived from satellite imagery) is crucial to quantify map accuracy and reliability. Since it is impractical to ground-truth every pixel or mapping unit in a large area, practitioners rely on *sampling strategies* to select a representative subset of locations for accuracy assessment ¹ ². Two fundamental approaches are **random point sampling** and **random area (polygon/cluster) sampling**. In random point sampling, individual point locations (often corresponding to single map pixels) are randomly selected for verification, whereas random area sampling selects spatial units such as blocks of pixels, segments, or polygons at random for validation. The choice of sampling design affects the statistical soundness of the accuracy assessment – including bias control, precision of accuracy estimates, and spatial representativeness ³ ². This review traces the historical development of random sampling in spatial analysis and remote sensing, examines how point vs. area-based random samples are used in practice, identifies criteria for good sampling design, compares random sampling with stratified, systematic, and cluster designs across different landscapes, and highlights open research questions. Peer-reviewed literature (e.g. *Remote Sensing of Environment*, *ISPRS J. Photogrammetry & RS*, *IEEE TGRS*) is emphasized, with references in APA format.

Historical Development of Random Sampling in Spatial Analysis

The roots of statistical sampling for spatial data can be traced to general survey sampling theory (e.g. Fisher, Cochran in the mid-20th century), but its formal adoption in remote sensing accuracy assessment began in the 1970s ⁴. Early land-use mapping projects recognized that some form of probability sampling was needed to estimate map accuracy with statistical confidence ¹. One pioneering study by **Hord and Brooner (1976)** provided mathematical criteria for determining the number of sample points needed for map accuracy evaluation ⁵. Around the same time, **Van Genderen and Lock (1977)** described a “simple statistical sampling procedure” for testing land-use map accuracy, arguing that a minimal number of field-check sites should be visited while still yielding statistically valid results ⁶ ⁷. These early works introduced the error matrix (confusion matrix) concept and underscored the impracticality of exhaustively checking every map parcel ¹. By the late 1970s, **stratified random sampling** had already been recommended as an optimal strategy for remote sensing studies so that even small or rare classes would be represented in the reference data ⁷. For example, Rudd (1971) and Zonneveld (1972) (as cited by Van Genderen & Lock) advocated stratifying by map category to ensure smaller land-use classes are sampled ⁷.

In the 1980s, the practice of accuracy assessment became more routine. **Story and Congalton (1986)** reinforced the use of error matrices and sampling considerations for land-cover classifications. By 1990, agencies and stakeholders often required maps to meet minimum accuracy standards (commonly 85% overall accuracy) ⁸, prompting more systematic assessment approaches. **Congalton (1991)** published a seminal review of classification accuracy assessment in *Remote Sensing of Environment* that synthesized best

practices up to that point ⁹. Congalton highlighted that the validity of the accuracy results hinges on the sampling approach: the sample of reference sites must be representative of the entire mapped area, otherwise the computed accuracy metrics are meaningless ¹⁰ ¹¹. He advocated a combination of stratification and random selection – typically, stratify by map class then sample reference points randomly within each class ¹². This approach ensured even small classes were evaluated while maintaining the unbiased nature of random sampling. A common rule of thumb emerged from this period: *aim for at least 50 validation samples per map class* ¹³ ¹⁴. This guideline, originally suggested by researchers like Hay (1979) and echoed by Congalton (1991), is still often followed to ensure sufficient within-class sample size for reliable error estimates ¹⁵ ¹⁴.

Through the 1990s and 2000s, statisticians refined sampling designs for accuracy assessment. **Stehman and Czaplewski (1998)** laid out rigorous principles for design-based inference in thematic map accuracy studies, framing accuracy assessment in three components: sampling design, response design, and analysis ¹⁶ ¹⁷. They and others (e.g. Stehman 1999; 2009) enumerated the pros and cons of basic designs – simple random, stratified random, systematic, and cluster sampling ¹⁸ ¹⁹ – providing the community with clearer guidance. By the 2010s, large-area mapping projects (e.g. national land cover inventories, global land-cover maps) adopted probability sampling protocols as *best practice*. For instance, the international good practice guidelines by Olofsson et al. (2014) recommended stratified random sampling and probability-based estimators for area and accuracy, rather than error-prone practices like using map pixel counts ²⁰. Yet, despite a half-century of research and published guidelines, many operational projects still deviated from rigorous designs. A recent literature review of 282 studies (1998–2017) found only about **54%** had employed an explicit probability sampling design for validation (most commonly stratified random) ², and only one-third of studies documented enough detail to be fully reproducible ²¹ ²². This underscores that while the science of random sampling in remote sensing is mature ²³, its practice is not yet universal – a gap discussed further in this review's conclusion.

Random Point vs. Random Area Sampling in Accuracy Assessment

Random point sampling refers to selecting point locations at random across the map for ground-truth validation. In raster classification accuracy assessments, this often equates to randomly choosing individual pixels ¹⁷. Each point/pixel's mapped class is compared to the true class determined from higher-quality reference data (field visit, high-resolution imagery, etc.). The point sampling approach is straightforward and treats every pixel equally, aligning with the assumption of most accuracy estimators that each sampled unit has the same area weight ²⁴. Random point sampling has been widely used in remote sensing because it is easy to implement (e.g. generating random coordinates) and supports simple design-based analysis (each sample an independent Bernoulli trial of correctness) ²⁵. However, pure random points can sometimes be suboptimal if the map's minimum mapping unit is larger than a pixel or if geolocation uncertainty exists – a single isolated pixel may not reliably represent a map class on the ground ²⁶. For this reason, some guidelines *discourage using single pixels as reference units* and instead recommend aggregating or using larger homogeneous units ²⁶. Despite this caveat, random pixel sampling remains very common (the majority of validation datasets in one review used individual pixels as the reference unit) ²⁷. Notably, Morales-Barquero et al. (2019) found that studies using pixel-level reference data tended to report slightly higher accuracies, potentially because matching a single pixel's class is easier or because small reference plots may misalign with coarse pixels ²⁸ ²⁹. This suggests random point sampling can sometimes overestimate accuracy if reference data are not well-aligned, but in general it provides an unbiased assessment when properly executed ²⁴.

Random area sampling involves selecting spatial areas (rather than points) at random for validation. The “area” could be a contiguous block of pixels, an image-object or segment, or a reference polygon (e.g. a field plot or an entire land-cover patch) ¹⁷ ³⁰. This strategy is often motivated by practical and methodological considerations. Practically, if conducting field surveys, it can be more efficient to validate a cluster of points in a compact area (reducing travel time) rather than truly random scattered points – thus leading to *cluster sampling*, a form of area-based sampling where entire clusters (areas) are chosen randomly ³¹. Methodologically, area units may better match the way a map was produced. For instance, in object-based image analysis (OBIA), the basic map units are segments/polygons rather than individual pixels ³⁰. In such cases, using whole polygons as the sampling unit (randomly choosing map objects to check) ensures the validation aligns with the mapping unit: each sampled object is labeled entirely correct or incorrect ¹⁷. Even in pixel-based maps, some accuracy assessments use *plot-level reference data*, where a small area (e.g. a 3x3 pixel block or a 1-hectare field plot) is surveyed and its dominant land-cover recorded ³² ²⁶. The mapped pixels falling within that plot can then be validated in aggregate. Random area sampling can thus help when a single pixel is too fine a unit to evaluate (due to mixed pixels or GPS error) and can provide additional information, such as spatial autocorrelation of errors within an area ³³ ³⁴.

However, using varying-size polygons or clusters as samples introduces complexities. One key issue is **unequal inclusion probabilities** – large areas cover more map pixels than small areas, yet if each area is sampled with equal probability, a large land-cover patch and a small patch are given the same weight in accuracy calculations. This can bias accuracy estimates ³⁵. To avoid this, researchers have developed estimators that weight each sampled polygon by its area or size when computing accuracy metrics ³⁵ ²⁴. *Radoux and Bogaert (2014)* demonstrated that not accounting for polygon size can misestimate overall accuracy and class accuracies, and they proposed area-weighted formulas to correct this ³⁵. Another consideration is that when an area is the sample unit, one must define the response design carefully: is the area “correct” only if *all* pixels in it are correctly classified, or if a majority are correct, or by some fuzzy overlap measure? Different studies have used different criteria ¹⁷ ³⁶. Typically, for polygon-based mapping, if the entire object is given one class on the map, the object is counted as correct if its predominant ground truth class matches the map label ³⁰. For cluster sampling of pixels, sometimes all pixels in the cluster are individually checked and an average agreement is computed, or the cluster is treated as multiple point samples but with a variance correction in analysis ³⁷. Despite these nuances, random area sampling strategies have proven *viable and often necessary* in many cases (e.g. assessing map accuracy over large forest regions with clustered field plots, or validating object-based classifications) ³⁰ ²⁵. Liu et al. (2005) compared point-based versus polygon-based accuracy assessment on the same imagery and found that **both yield valid results, but point sampling gave slightly higher precision** in their case ²⁵. The polygon sample method showed a bit more variability, possibly due to mixed-class polygons or the all-or-nothing scoring of a polygon’s correctness ²⁵. Nonetheless, polygon sampling is indispensable for validating maps whose legends or minimum mapping units are defined over areas larger than a single pixel (such as mapping forest stands or agricultural fields). Modern accuracy assessments thus often combine approaches: for example, a national forest map might use stratified random *cluster sampling*, where a number of Landsat-sized blocks are randomly chosen and within each, a grid of sub-sample points is validated ³² ³⁸. This hybrid ensures broad coverage while leveraging field effort efficiently. In summary, random point and random area sampling are both widely used – the former offering simplicity and direct unbiased estimation, and the latter offering better alignment with mapping units and field logistics, at the cost of more complex analysis.

Criteria for an Effective Sampling Strategy

Designing a good sampling strategy for accuracy validation requires balancing statistical rigor with practical constraints. Key criteria identified in the literature include:

- **Probability-Based Selection:** A fundamental principle is to employ a probability sampling design – *every unit (pixel or area) in the map should have a known, non-zero chance of selection* ³⁹ ². This ensures the sample is statistically representative of the map population, allowing unbiased inference about map accuracy. Non-probability sampling (e.g. hand-picking “representative” sites or only sampling easily accessible areas) can introduce unknown bias ⁴⁰. Unfortunately, studies show that many accuracy assessments still rely on ad-hoc or purposive sampling, which undermines confidence in the reported accuracies ⁴⁰ ²². Good practice demands randomization in selecting validation points, possibly stratified or systematic, but never purely convenience-based ⁴⁰.
- **Spatial Coverage and Representativeness:** The sample should be spatially distributed to cover the map area and all its variation. A poor design might, by chance, cluster many samples in one region while leaving other areas unvisited (especially with small sample sizes and purely random points). Thus, designs that enforce spatial spread – such as stratifying by geographic subregions or using systematic grids – can improve coverage ⁴¹ ³³. In land-cover maps, *covering all map classes* is critical: a good sample allocates samples to capture both dominant and rare classes ⁴² ⁴³. Stratified random sampling by class is often recommended precisely to guarantee even the smallest classes are represented in the reference data ⁷ ⁴². For example, the U.S. Bureau of Land Management suggests a post-classification stratified random sample so that **all** land-cover categories get sampled at least a minimum number of times ⁴³. Ensuring broad spatial and thematic coverage makes the accuracy results more reliable for the entire map domain.
- **Sufficient Sample Size:** The reliability of accuracy statistics (e.g. overall accuracy, per-class accuracy) depends on having an adequate number of sample observations. Too small a sample yields large confidence intervals and unstable estimates. As noted, a common benchmark is at least 50 samples for each class of interest ¹⁴ ⁴⁴. This rule, while not derived from a specific formula, has been adopted to balance effort and precision, and it aligns with analysis showing precision gains level off beyond ~50 samples per class in many cases. Larger areas or more complex maps (many classes) generally require larger total sample sizes ⁴⁵. Several authors (e.g. Stehman, 1997; Curran & Hay, 1986) have provided equations to calculate sample size given desired confidence levels, but these often require an initial guess of accuracy. In practice, an iterative approach may be used: allocate a baseline (50 per class), then adjust upward for highly critical classes or to narrow margins of error. *Balanced accuracy* (giving each class equal samples) versus *proportional allocation* (sampling classes in proportion to their area) is another consideration – each has implications for precision and the weighting of error matrix results ¹⁵ ⁴³. A good design explicitly considers sample size allocation and, when reporting results, provides the sample counts and perhaps confidence intervals ⁴⁶ ².
- **Bias Control and Independence:** A sound strategy avoids biases such as *sampling the training data* or preferentially sampling certain map areas. It is imperative that **validation data are independent of training data** used to create the map ⁴⁷. If the same locations were used both to train the classifier and to assess accuracy, the accuracy results will be optimistically biased (sometimes dramatically so). Thus, many protocols either (a) reserve completely independent field data for accuracy assessment, or (b) if using one dataset for both training and testing, ensure a clear

separation (e.g. spatially or temporally) to simulate independence⁴⁷. Another bias issue is positional bias – e.g. systematically choosing sample points near roads for convenience will bias accuracy high for classes found near roads (and leave remote areas untested). Truly random or systematic designs help mitigate this by not correlating selection with accessibility. Additionally, if cluster sampling is used, one must account for *intra-cluster correlation*; otherwise, treating clustered points as independent inflates the effective sample size and can mislead uncertainty estimates⁴⁸. Good practice is to apply appropriate variance estimation techniques (e.g. a cluster-adjusted variance estimator or bootstrapping) to reflect the actual information content of clustered samples (Stehman, 2014). In summary, a good design is one that yields unbiased *and* independent samples for truthful accuracy estimation.

- **Consistency with Map Resolution and Minimum Mapping Unit:** The sampling unit should be chosen in accordance with the map's characteristics. If the map's minimum mapping unit (MMU) is, say, 1 hectare (meaning the map generalization does not show patches smaller than 1 ha), then using a 30 m pixel as a sample might be problematic – a single pixel could be mislabeled due to map generalization, even though the larger MMU block is correctly labeled. In such cases, using a **block or polygon** of area equal to the MMU for validation can be more appropriate²⁶. The U.S. National Vegetation Classification accuracy assessment guidelines, for example, emphasize matching the reference sample unit to the map's MMU (whether that's a cluster of pixels or a polygon)²⁶. This ensures that the comparison between map and reference is fair. If using points when the map is polygon-based, one must carefully decide how to handle points that fall near class boundaries or within mixed pixels. Clear **response design** rules (e.g. only sample points in interior of polygons, or use majority-area label in a plot) are part of a good strategy^{17 36}. In short, the spatial support of the reference data should correspond to the mapped spatial unit to avoid inconsistencies.
- **Transparency and Reproducibility:** A sometimes overlooked but vital criterion is how well the sampling strategy is documented and reproducible. Accuracy assessments should be reported with a clear description of the sampling design (e.g. "We used stratified random sampling with strata = map class, allocating 30 samples per class... and here is how reference labels were obtained...")²². Without this, users of the map or readers of the study cannot fully judge the credibility of the accuracy figures²². Reproducibility also allows future researchers to repeat the assessment or apply the same protocol elsewhere. Good practice includes publishing or sharing the actual list of reference sample locations and their reference labels whenever possible^{49 50}. Recent literature has noted a "lack of transparency" in many studies – with omissions about how samples were collected, what constituted a "correct" reference, etc., making results hard to verify²². To meet this criterion, one should adhere to reporting standards (such as those suggested by Stehman & Foody, 2019) and possibly follow established accuracy assessment **protocols** (e.g. the CEOS Land Product Validation sub-group's guidelines). In summary, the quality of a sampling strategy is not just in its design, but also in how well it's executed and communicated.

By meeting the above criteria – probability-based, well-distributed, adequately sized, unbiased and independent, appropriate to the map scale, and transparently reported – an accuracy assessment can be considered robust. These criteria align with the concept of an accuracy assessment that is "rigorous, informative, and honest"⁵¹, providing users with confidence in the reported accuracy metrics.

Comparison of Sampling Methods for Different Landscapes

Various sampling designs have been applied to remote sensing validation, each with strengths and weaknesses. Here we compare simple random sampling with other common designs – stratified, systematic, and cluster sampling – and discuss their performance in different landscape contexts (e.g. urban vs. forested vs. agricultural environments).

- **Simple Random Sampling (SRS):** In simple random sampling, every map unit (pixel or area) has an equal chance of being selected. The design is straightforward and statistically *unbiased*, making analysis simple (standard error formulas assume independent random samples) ⁵² ³⁶. The advantage of SRS is its conceptual simplicity and lack of need for ancillary information (no strata or grid needed). In a homogeneous landscape (e.g. a vast forest with one dominant cover type), pure random sampling can adequately capture accuracy with proper sample size – though many samples will fall in the dominant class. However, in heterogeneous landscapes, SRS may *by chance* miss some small classes or leave large spatial gaps. For instance, an urban area with many small land-cover patches might not have any samples from a rare class like “water” if samples are allocated purely randomly by area proportion. SRS also does not guarantee spatial balance; one corner of the map might contain many random points while another has few. Despite these issues, SRS remains a baseline; it is often used when no strong ancillary stratification is available or in combination with stratification (e.g. within each class) ¹². One empirical finding is that SRS tends to have higher variance in accuracy estimates compared to more structured designs if the variable of interest (classification error) is spatially autocorrelated ⁴². In other words, if errors cluster in space, a purely random sample might, by luck, hit an error cluster or miss it entirely, causing more variability. Designs that spread samples out (like systematic) can mitigate that. Nevertheless, SRS is unbiased and easy to implement and remains a core component of many designs (often as the method of selection within strata or clusters).
- **Stratified Random Sampling:** Stratified sampling divides the map into distinct *strata* and conducts random sampling within each stratum. Common stratifications are by **map class** (allocate samples to each land-cover class) or by geographic zones (e.g. eco-regions, administrative regions) ⁵³ ⁵⁴. The major benefit of stratified random sampling is improved representation of all parts of the map. As noted, stratifying by class ensures even minor classes are evaluated (at least some samples per class) ⁷ ⁴². This often leads to **higher precision** in estimating per-class accuracies and can also improve overall accuracy estimates if done with optimal allocation ²⁵. Liu et al. (2005) found that in their tests, **stratified sampling produced more precise accuracy estimates than simple random or purely systematic designs** because it forced coverage of all land-cover types ⁴². Stratification can also be by region: for large areas, one might stratify the map into urban, agricultural, and forest zones, or by continents in a global map, to ensure each region is sampled. Stratified designs do require knowledge of the strata (e.g. one needs a preliminary map to stratify by class, or ancillary data to stratify by eco-region) ⁴³. They also require using appropriate weighted estimators when combining strata results – e.g. if equal samples per class are taken, the **analysis** must weight accuracy contributions by class area to get an unbiased overall accuracy ³⁹ ⁵⁵. Congalton (1991) suggested a two-phase approach: use stratified sampling during map production (e.g. collecting training data per class), then after classification, take an independent stratified random sample for accuracy assessment ¹². This underscores that training and validation can both benefit from stratification but must use separate points. In practice, stratified random sampling is **widely regarded as the preferred method** for land-cover map accuracy assessment ⁴³. Most high-profile

mapping projects (e.g. NLCD in the US, CORINE in Europe) use a stratified random design by class or land-cover strata. The only limitations are: if a class occupies a tiny area, how to obtain samples (sometimes via oversampling that class beyond its proportion); and if the number of strata is large, total sample size must be large to give each stratum enough samples. In diverse landscapes, stratified sampling is highly effective – e.g. in a mixed urban/agricultural region, one can stratify by land-cover type (urban, crop, water, forest, etc.) and ensure even small urban green spaces or water bodies get assessed. This yields a more informative and stable accuracy profile across classes ⁴². Stratification can also incorporate spatial stratifiers: e.g. **stratified systematic unaligned sampling** breaks the area into grid cells and randomly picks one site per cell, effectively stratifying by location to enforce spread ³⁴. Overall, stratified random designs combine the unbiased nature of random sampling with the representativeness of structured allocation – making it a powerful approach in remote sensing validation.

- **Systematic Sampling:** In systematic sampling, sample units are selected at regular intervals over space (e.g. every 10 km, or on a fixed grid) ³³. This approach ensures a very uniform spatial coverage and is logically easy to implement in the field (sampling points can be pre-determined in a grid). The advantage is that no part of the map is too far from a sample – it inherently achieves good spatial spread ⁵⁶. Systematic sampling can be *more precise* than random sampling if the phenomenon of classification error is spatially autocorrelated (which it often is; errors might be regionally clustered). By evenly covering the area, systematic designs avoid clumping of samples that might occur in random draws. However, the **drawback** is that systematic sampling is not a true probability sample in a strict sense (only the starting position is random, thereafter positions are fixed). This complicates formal uncertainty estimation because there is no straightforward formula for variance with one systematic realization (one often assumes it behaves like a random sample for approximation, or uses spatial models). More seriously, if there is any hidden *periodicity or pattern* in the landscape aligning with the systematic interval, it can bias results ⁵⁷. An infamous caution is if the sampling interval coincides with a regular pattern on the map (e.g. a 5-km grid sampling in a region where land-use changes every 5 km in a repeating way), the sample could over- or under-represent certain classes ⁵⁷. Fenstermaker (1991) noted that systematic transects could give poor accuracy estimates if the landscape has periodic variation that syncs with the transect spacing ⁵⁸. One variant that mitigates this is **stratified systematic unaligned sampling**, where the area is divided into grid cells (strata) and within each cell, a random point is chosen ⁴¹. This yields one sample per grid cell, avoiding alignment with any single periodic structure, and is a recommended compromise (used in some forestry surveys and satellite product validations) ³⁴. In terms of landscape context: for a relatively uniform environment (e.g. a large forest or desert), systematic sampling works well and can give very precise overall accuracy estimates because it captures the smooth spatial trends. For highly heterogeneous areas (e.g. urban with patchy land use), systematic sampling ensures coverage, but one must be careful with scale – a fine grid may be needed to capture small patches, and the grid origin should be randomized to avoid bias ³³ ³⁴. Systematic sampling is also common in high-resolution image mapping where a grid (perhaps every 100th pixel) is sampled to get a quick accuracy read. Overall, systematic designs are favored for their logistical simplicity and spatial balance, but analysts must assume (or demonstrate) that no harmful spatial periodicity exists in the map errors. If such an assumption holds, systematic samples can yield lower standard error of accuracy estimates than an equivalently sized random sample in many cases ³³ ⁴¹.

- **Cluster Sampling:** Cluster sampling involves selecting groups of pixels or map units (clusters) as the sampling unit, rather than individual points scattered across the map ³¹ ⁴⁸. In a two-stage cluster design, one might first randomly select, say, 30 map cells (primary sampling units), and then within each cell, randomly sample 10 pixels for verification (secondary units). The chief motivation for cluster sampling in remote sensing is **field efficiency** ³¹. If ground validation is needed, it is much easier to visit a few areas and check many points within those areas than to travel to 300 randomly located points spread all over. Cluster sampling can drastically cut travel time and cost when the study area is large or difficult to access. For example, a national forest inventory might randomly pick certain 10×10 km squares (clusters) and validate all accessible points or plots inside them, instead of a national random sample of points that could be on mountaintops or deep wilderness. Cluster sampling is also useful when the map accuracy needs to be reported at a *block or polygon level*, not just per pixel. However, the trade-off is statistical: points within a cluster are usually more similar to each other (positive spatial autocorrelation) – if one point in a cluster is misclassified, its neighbors might be misclassified too. This *intra-cluster correlation* means you get less independent information from 20 points in one cluster than from 20 points spread randomly. In essence, cluster sampling often yields a higher variance (larger standard error) for accuracy estimates compared to a simple random sample of the same size ⁴⁸. Stehman (1992) showed that if cluster size is large, the effective sample size reduces (because of redundancy within clusters), so one must increase total sample size or number of clusters to compensate. Another requirement is to use proper formulas for accuracy and variance under cluster sampling – treating each point as independent (as if SRS) will underestimate the uncertainty ⁴⁸. Fortunately, statistical techniques like the *jackknife* or *Taylor series expansion* can estimate variance from cluster samples (e.g. see Stehman, 1997; Wolter, 1985 for methods), and many accuracy assessment studies have employed them ⁴⁸ ³¹. In practice, cluster sampling is commonly combined with stratification: e.g. stratify the map by land-cover class, then cluster sample within each class to ensure both the benefits of stratification and logistical efficiency. One study noted challenges with such combination – ensuring enough clusters per class and analytical complexity – but it is feasible (Stehman, 2009) ⁵⁹. By landscape type, cluster sampling is particularly useful in **remote or extensive areas** like forests, rangelands, or wetlands where travel is difficult. For instance, in a forest accuracy assessment one might sample a limited number of forest stands (clusters) and within each stand lay out several plots to verify map labels. In contrast, in a compact urban environment, cluster sampling might not be needed (travel distances are short) and might even be undesirable if it accidentally concentrates samples in a few neighborhoods. In agriculture, cluster designs might sample a handful of farms and check many fields within each – again trading some statistical efficiency for practical ease. A second motivation for cluster sampling highlighted by Stehman (1992) is when the *map's unit of interest is larger than a pixel* ³¹. For example, if a map is used to estimate crop area, cluster (area) accuracy metrics might better capture whether an entire farm or field is correctly classified rather than isolated pixels. In summary, cluster sampling is a useful strategy in large-area and field-based validation contexts; it must be accompanied by proper analytic adjustments for clustering. Its strengths are cost-effectiveness and ability to incorporate area units, while its weaknesses are lower precision per sample and more complex error analysis ⁴⁸.

Landscape considerations: Different environments can influence which sampling design is optimal:

- *Urban landscapes* – highly heterogeneous with many small patches (parks, buildings, roads). **Stratified random** sampling by land-cover type is often favored in urban studies ⁴², to ensure even small classes (e.g. water bodies, bare soil) are included. A purely random or coarse systematic

sample might miss some tiny classes entirely or sample them too sparsely to evaluate class-specific accuracy. **Systematic sampling** on a fine grid can also work, as it guarantees a spread across city neighborhoods; however, care must be taken that the grid interval isn't too large (which could skip over narrow features like streets or rivers). If using systematic transects or grids in urban areas, an *unaligned* approach (random start in each grid cell) is recommended to avoid coinciding with city block layouts or other regular patterns ⁴¹. Clustering is usually not needed in cities because travel within a city is not as onerous as in wilderness, and clustering could cause oversampling of particular districts. Thus, urban accuracy assessments often end up using stratified random points, sometimes supplemented by manual checking of certain critical small features (which is a form of purposive sampling, albeit not ideal). For example, a study of urban land cover might stratify by land use zones or by class (built-up, vegetation, water, etc.) and randomly sample points in each, ensuring representation of rare urban classes like wetlands or industrial areas.

- *Forested or natural landscapes* – often expansive and homogeneous over large areas, but with some rare cover types (water, burn scars, etc.). For large forest regions, **cluster sampling** can be very efficient: field crews can be sent to a few randomly picked forest sites (clusters) and validate many points or plots there. This was the approach in the US National Land Cover Dataset accuracy assessments, where blocks of pixels were photo-interpreted rather than totally random points, due to field logistics ³¹. **Stratification** is still useful: one might stratify by forest type or by ecozone (e.g. lowland vs upland forest) and then cluster-sample within each. If the forest map has one dominant class (forest) and very small proportions of others (say wetlands, clearings), stratification by class is crucial to capture enough non-forest points for analysis ⁴³. **Systematic sampling** can be useful if remote sensing analysts use moderate resolution imagery and can systematically sample points for visual verification with higher-resolution imagery (e.g. a regular grid of points checked on Google Earth). This has been done in global forest change maps (Hansen et al. 2013 used a systematic sample of reference tiles). The risk of bias from systematic sampling in natural landscapes is low unless there's a regular pattern (which nature rarely has, except maybe geology). So systematic or stratified-systematic designs are quite effective for broad-area environmental maps ⁴¹. In dense forest, a purely random sample might leave large unsampled gaps, so a grid ensures even coverage. Overall, for forests, a mix of stratification (to ensure minor classes like non-forest cover are included) and clustering (to manage field effort) is common.
- *Agricultural landscapes* – typically semi-regular patterns (fields, crop rotations) and often a mix of large uniform fields and smallholder patches depending on region. **Stratified sampling** by crop type can ensure that each crop's area is assessed, which is important if some crops are rare or if certain crop classes are much less accurately mapped than others. Agricultural maps also may warrant stratification by region if farming practices differ by region. A potential issue in agriculture is that fields often follow a grid or topographic pattern – e.g. in parts of the U.S., fields are laid out in square mile sections. A naive systematic sample aligned with that same grid could either consistently hit field centers or boundaries, introducing bias (if, say, boundaries tend to be mixed pixels and often misclassified). To avoid this, an *unaligned systematic* or random approach is preferable. Also, crops change every year, so **temporal considerations** come in: an accuracy assessment might stratify by time or phenology (ensuring samples in early and late season fields). **Cluster sampling** might be useful if ground surveys are done by selecting certain farms and validating multiple fields there. But one has to be cautious: if one farm uses unique practices affecting classification accuracy, cluster sampling many fields in that one farm could misrepresent the general accuracy. Ideally, many small clusters across different areas would be chosen rather than a few large clusters in agriculture.

Simple random sampling in agricultural areas can perform quite well if fields are large relative to the pixel size – randomly chosen points are likely to fall well inside homogeneous fields, yielding clear reference labels. But if fields are small or irregular, a pure random sample might include many mixed pixels near field boundaries, potentially underestimating accuracy (unless the response design accounts for mixed pixels). Thus, some studies use stratified random sampling that differentiates between interior-of-field and edge-of-field strata to evaluate map errors in each. In summary, agricultural map validation often employs stratified random designs (by crop type) or systematic grid sampling for even coverage, with possible clustering if ground visits are involved.

Overall, no single design is best for all situations – often a *hybrid approach* is optimal. For example, a **two-stage design** might first stratify the map into zones (or classes), then within each use a systematic sample or cluster sample. Stehman (2009) emphasizes that the choice depends on map objectives and practical constraints, recommending that designers weigh the increase in precision against complexity and cost ⁶⁰ ₆₁. The table below summarizes the key points of each sampling method in context:

Sampling Design	Strengths (for accuracy assessment)	Limitations	Best Uses / Context
<i>Simple Random</i>	Unbiased; simple to implement and analyze ⁵² .	Could miss rare classes; not spatially balanced (higher variance) ⁴² .	Baseline method; use when no info to stratify or in fairly uniform areas.
<i>Stratified Random</i>	Ensures all strata (e.g. classes) are sampled – higher precision per class ⁴² ; can allocate optimally to improve efficiency.	Requires strata info (map or ancillary); must weight results by area ³⁹ .	Heterogeneous maps (urban, mixed land-cover); any case where some classes or regions are small but important ⁷ .
<i>Systematic (Grid)</i>	Excellent spatial coverage – spreads samples evenly ⁴¹ ; often more precise if spatial autocorrelation in errors.	Not strictly random (variance estimation is tricky); risk of bias if periodic spatial patterns align ⁵⁷ .	Large-area mapping (e.g. country-scale) for overall accuracy; when field access is not an issue and want even coverage (e.g. using photo-interpretation).
<i>Cluster (Area) Sample</i>	Cost-effective for field data – many samples in one trip ³¹ ; can match map MMU by using area units.	Lower effective sample size (intraclass correlation) – higher SEs ⁴⁸ ; analysis more complex (need cluster-adjusted calculations).	Huge or hard-to-access areas (forests, global maps); object-based maps (sampling whole objects); situations requiring ground teams to minimize travel.

(Sources: Congalton 1991; Stehman 1992; Stehman & Czaplewski 1998; Liu et al. 2005; Olofsson et al. 2014)

It's worth noting that in practice, many accuracy assessments blend these methods. For example, a project might use **stratified random sampling by class, with one cluster of 5 points per sample** (so-called *stratified two-stage cluster sampling*). Another might use a **stratified systematic** approach (randomly select

one point per map tile). Each combination inherits pros and cons of its components. The unifying theme is using randomness to avoid bias while possibly imposing structure to ensure coverage. The literature strongly suggests that *any* of these probability-based designs is preferable to non-probabilistic methods (like using only “easy” reference sites), which can greatly mislead accuracy results ⁴⁰ ²².

Open Research Questions and Gaps in the Literature

Despite substantial progress in sampling strategies for map validation, several open questions and challenges remain:

- **Closing the Gap Between Theory and Practice:** As noted, only about half of remote sensing studies (1998–2017) explicitly used probability sampling for validation, and even fewer fully reported their methods ²¹ ²². Many maps are still validated with ad-hoc procedures that do not meet the rigorous criteria discussed. This gap raises the question: *How can the community encourage and enforce better sampling practices?* Recent papers call for greater transparency and for journals/editors to demand that authors report sampling designs and error matrices ²² ⁶². Some efforts, like the introduction of standardized validation protocols (e.g. ISO standards or CEOS best practice documents), aim to make good sampling design more accessible. However, uptake is slow. Research could explore incentives or tools (perhaps easy-to-use software for sample design) to help practitioners implement proper sampling. Another idea is developing automated auditing of published accuracy results to flag if the design likely had bias (Castilla 2019 speculated many published accuracies would fail scrutiny) ²².
- **Imperfect and Evolving Reference Data:** Traditional accuracy assessment assumes reference labels are 100% correct (“ground truth”). In reality, reference data (whether from field observation or higher-res imagery) can have errors and uncertainty. **Foody (2015)** and others have pointed out that *imperfect reference data merit greater consideration*, because validation samples that are themselves mislabeled will bias accuracy estimates ⁶³ ⁵¹. Research questions remain on how to design sampling or analysis to account for reference uncertainty – for example, should one model reference error or use multiple reference observers to estimate reliability? Some recent work uses confusion matrices for both map and reference (the latter derived from inter-observer studies) to adjust accuracy estimates (pers. comm. Olofsson et al. 2021). As the volume of reference data grows with citizen science and automated labeling, dealing with reference uncertainty statistically is an open challenge.
- **Scaling Sampling to Big Data (Space and Time):** With modern Earth observation, land-cover maps are produced at *global scales* and updated frequently (e.g. annual or even near-real-time change maps). Validating such products stretches traditional sampling approaches. New challenges arise in mapping “extensively in space and intensively in time” ⁶⁴. How to sample a global map to make inferences at global, continental, and national scales simultaneously is one question. One approach is multi-stage stratified sampling (as used in global forest assessments – dividing the globe into regions, then sub-regions, etc., and sampling hierarchically). Another aspect is temporal sampling: for a yearly land cover product, does one validate every year separately (requiring new samples each year), or is there a way to leverage samples across time? A *spatiotemporal sampling design* could, for example, track a panel of sample locations over time to assess change accuracy, but that introduces temporal correlation issues. Research is ongoing into efficient strategies for accuracy assessment of time-series maps or near-real-time classifications. The technology of high-frequency mapping (e.g.

Sentinel-1 and -2 providing bi-weekly land cover updates) has outpaced our validation capacity – it's infeasible to send field teams constantly. So, investigating use of new reference sources (like daily high-res cubesats or crowdsourced data) in a statistically sound way is an open problem.

- **Spatial Autocorrelation and New Estimation Techniques:** Classical accuracy estimation assumes samples are independent. In reality, map errors often exhibit spatial autocorrelation (e.g. an entire region might be misclassified due to, say, clouds or a classification confusion). Traditional variance formulas may underestimate the uncertainty if positive autocorrelation is present. Conversely, a well-spread sample (systematic) might benefit from autocorrelation by effectively sampling more of the variance. *Geostatistical techniques* and *spatial simulation* could potentially provide more accurate confidence intervals by explicitly modeling spatial error correlation. Some recent studies have proposed using variogram modeling of error or spatial bootstrap methods (e.g. moving block bootstrap) to assess map accuracy uncertainty under spatial dependence (e.g. Zhao et al. 2019 in *IEEE TGRS*). This is a frontier where statistical theory (design-based vs model-based inference) intersects with practical needs. The debate continues on how to combine the rigor of design-based (randomization) approaches with spatial models to improve precision without sacrificing unbiasedness ⁶⁰ ⁶¹.
- **Integration of Stratification Criteria:** While class-based stratification is common, there is scope to explore **alternative stratifiers** to improve efficiency. For instance, stratifying by predicted local accuracy (from classifier outputs like confidence or probabilistic maps) – i.e. allocate more samples to areas suspected to have lower accuracy. This ventures into *adaptive sampling* or *model-informed sampling*, which blurs design- and model-based inference but could yield more information where it's most needed. Some researchers have looked at using ancillary variables (e.g. NDVI variance, elevation zones) for stratification when class stratification alone is insufficient (Stehman 2014 mentions stratifying in space and time for burned area product validation ¹⁸ ¹⁹). Open questions include how to optimally stratify in multi-dimensional ways (classes, regions, time periods) and how to allocate samples among these stratifications.
- **Automation and Reproducible Workflows:** Another gap is the lack of easily reusable workflows for probability sampling in GIS software. Many practitioners still find it easier to drop pins on a map by eyeball than to program a stratified random selection. Tools and libraries are emerging (e.g. some QGIS/ArcGIS plugins or the R package *mapaccuracy*), but more could be done. Ensuring these tools also output clear documentation (so that reporting is improved) is a practical area of development rather than pure research.
- **Validation of New Mapping Paradigms:** As geospatial analysis evolves (e.g. deep learning classification, crowdsourced mapping, continuous field maps instead of discrete classes), validation strategies may need to adapt. For example, image segmentation methods yield objects – how to sample objects in a way that accounts for both their thematic accuracy and their boundary accuracy is an open question (object accuracy involves not just the label but also spatial delineation). Some recent work (Radoux et al. 2016; Persello et al. 2019) has looked at *object-based accuracy indices*. These require new sampling considerations, such as sampling object boundaries to assess boundary precision. Similarly, for *continuous variables* (like biomass or impervious surface fraction maps), random sampling is straightforward for estimating RMSE, but capturing spatially rare phenomena (like very high biomass stands) might benefit from stratification.

In conclusion, the field of accuracy assessment has come a long way – from early ad hoc checks to sophisticated sampling designs rooted in statistical theory. Random point and area sampling strategies form the backbone of most modern validation efforts, enabling quantitative statements about map quality. A long historical trajectory from the 1970s to today's best practices shows continual refinement toward more rigorous and transparent methods. Random sampling (in its various flavors: simple, stratified, systematic, cluster) remains essential for unbiased accuracy estimation, each method offering tools to handle different landscape configurations and practical constraints. A “good” sampling strategy today is characterized by probabilistic selection, solid coverage of map variation, sufficient sample size for stability, and careful attention to bias avoidance and documentation. While consensus on these principles is strong in the literature ⁶⁵ ⁶⁶, real-world practice doesn't always align, highlighting a need for ongoing education and perhaps enforcement of standards. Emerging challenges such as massive data scales, temporal map sequences, and imperfect reference information ensure that research on sampling strategies continues to be relevant. By tackling these open questions, the remote sensing community can improve the credibility and utility of the ever-growing array of geospatial data products. As one recent paper put it, only with *rigorous and honest* accuracy assessments can we turn our maps from “pretty pictures” into reliable scientific tools ⁶⁷ ⁶⁸.

References (APA)

- Congalton, R. G. (1991). *A review of assessing the accuracy of classifications of remotely sensed data*. **Remote Sensing of Environment**, **37**(1), 35–46. doi:10.1016/0034-4257(91)90048-B ⁹ ¹²
- Congalton, R. G., & Green, K. (1999). *Assessing the accuracy of remotely sensed data: Principles and practices*. Boca Raton: Lewis Publishers. ⁴ ¹⁴
- Hord, R. M., & Brooner, W. (1976). Land-use map accuracy criteria. **Photogrammetric Engineering and Remote Sensing**, **42**(5), 671–677. ⁵
- Van Genderen, J. L., & Lock, B. F. (1977). *Testing land-use map accuracy*. **Photogrammetric Engineering and Remote Sensing**, **43**(9), 1135–1137. ⁶ ⁷
- Story, M., & Congalton, R. (1986). Accuracy assessment: a user's perspective. **Photogramm. Eng. Remote Sens**, **52**(3), 397–399. (Discusses error matrices and sampling strategies for classification accuracy).
- Hay, A. M. (1979). Sampling designs to test land-use map accuracy. **Photogrammetric Engineering and Remote Sensing**, **45**, 529–533. (Proposed the often-cited 50 samples per class rule). ¹⁵
- Stehman, S. V. (1992). Comparison of systematic and random sampling for estimating the accuracy of maps generated from remotely sensed data. **Photogrammetric Engineering and Remote Sensing**, **58**(9), 1343–1350. (Found systematic sampling can be more precise but cautioned periodicity issues) ⁶⁹ ⁴¹.
- Stehman, S. V., & Czaplewski, R. L. (1998). Design and analysis for thematic map accuracy assessment: fundamental principles. **Remote Sensing of Environment**, **64**(3), 331–344. (Key reference outlining sampling design, response design, analysis framework) ¹⁶ ¹⁷.

- Stehman, S. V. (1999). Basic probability sampling designs for thematic map accuracy assessment. **International Journal of Remote Sensing**, **20**(12), 2423–2441. doi:10.1080/014311699212123 (Review of simple, stratified, cluster designs and their estimators).
- Stehman, S. V. (2009). Sampling designs for accuracy assessment of land cover. **International Journal of Remote Sensing**, **30**(20), 5243–5272. doi:10.1080/01431160903131000 (Comprehensive review of design options and recommendations) [60](#) [61](#).
- Brus, D. J., et al. (2011). Sampling for validation of digital soil maps. **European Journal of Soil Science**, **62**(3), 394–407. doi:10.1111/j.1365-2389.2011.01364.x (Evaluated simple, stratified, systematic, cluster, two-stage designs for soil map accuracy).
- Liu, X., He, C., Pan, Y., & Yang, M. (2005, July). Accuracy assessment of thematic classification based on point and polygon sampling units. *Proceedings of IGARSS 2005* (Vol. 6, pp. 4310–4313). IEEE. doi: 10.1109/IGARSS.2005.1525746 [25](#).
- Olofsson, P., Foody, G. M., Stehman, S. V., & Woodcock, C. E. (2013). Making better use of accuracy data in land change studies: Estimating accuracy and area and quantifying uncertainty using stratified estimation. **Remote Sensing of Environment**, **129**, 122–131. doi:10.1016/j.rse.2012.10.031.
- Olofsson, P., et al. (2014). Good practices for estimating area and assessing accuracy of land change. **Remote Sensing of Environment**, **148**, 42–57. doi:10.1016/j.rse.2014.02.015 (Guidelines endorsed by GOFC-GOLD, emphasizes probability sampling, stratified estimators) [20](#).
- Radoux, J., & Bogaert, P. (2014). Accounting for the area of polygon sampling units for the prediction of accuracy assessment indices. **Remote Sensing of Environment**, **142**, 9–19. doi:10.1016/j.rse.2013.10.030 (Introduced area-weighted accuracy estimators for polygon samples) [35](#) [24](#).
- Foody, G. M. (2015). Valuing map validation: The need for rigorous area estimation in map accuracy assessment. **Ecological Informatics**, **25**, 82–87. doi:10.1016/j.ecoinf.2014.10.007 (Discusses issues like imperfect reference data, importance of unbiased area estimates).
- Morales-Barquero, L., Lyons, M. B., Phinn, S. R., & Roelfsema, C. M. (2019). Trends in remote sensing accuracy assessment approaches in the context of natural resources. **Remote Sensing**, **11**(19), 2305. doi:10.3390/rs11192305 (Literature review showing only ~54% studies used probability sampling; calls for better reporting) [2](#) [22](#).
- Stehman, S. V., & Foody, G. M. (2019). Key issues in rigorous accuracy assessment of land cover products. **Remote Sensing of Environment**, **231**, 111199. doi:10.1016/j.rse.2019.05.018 (Summarizes 50 years of accuracy assessment, emphasizes design-based inference, future challenges) [63](#) [64](#).
- U.S. Bureau of Land Management (2013). *Accuracy Assessment Guidance* (Tech. Memo 2013-107). (Practical guidance document recommending stratified random designs, sample size rules, matching sample unit to MMU) [26](#) [43](#).

1 5 6 7 Testing Land-Use Map Accuracy

https://www.asprs.org/wp-content/uploads/pers/1977journal/sep/1977_sep_1135-1137.pdf

2 21 22 27 28 29 40 46 49 50 62 66 Trends in Remote Sensing Accuracy Assessment Approaches in the Context of Natural Resources

<https://www.mdpi.com/2072-4292/11/19/2305>

3 26 43 44 45 47 blm.gov

https://www.blm.gov/sites/blm.gov/files/uploads/IM2013-111_att5.pdf

4 13 15 18 19 20 23 51 63 64 65 67 68 Key issues in rigorous accuracy assessment of land cover products - ScienceDirect

<https://www.sciencedirect.com/science/article/abs/pii/S0034425719302111>

8 (PDF) Accuracy assessment and validation of remotely sensed and ...

<https://www.researchgate.net/publication/>

220040678_Accuracy_assessment_and_validation_of_remotely_sensed_and_other_spatial_information

9 10 11 12 14 32 33 34 38 41 53 54 56 57 58 69 [PDF] A Review of Assessing the Accuracy of Classifications of Remotely ...

<https://pure.iiasa.ac.at/id/file/198643>

16 17 24 30 35 36 39 52 55 Good Practices for Object-Based Accuracy Assessment | MDPI

<https://www.mdpi.com/2072-4292/9/7/646>

25 42 Accuracy assessment of thematic classification based on point and polygon sampling units | Request PDF

<https://www.researchgate.net/publication/>

4183487_Accuracy_assessment_of_thematic_classification_based_on_point_and_polygon_sampling_units

31 The SAGE Handbook of Remote Sensing

https://sk.sagepub.com/hnbnk/edvol/download/hdbk_remotesense/chpt/accuracy-assessment.pdf

37 Estimating standard errors of accuracy assessment statistics under ...

<https://www.sciencedirect.com/science/article/pii/S0034425796001769>

48 An Adaptive Sampling Design for Estimation of Thresher Shark ...

<https://www.researchgate.net/publication/>

267876768_An_Adaptive_Sampling_Design_for_Estimation_of_Thresher_Shark_CatchEffort_in_a_California_Recreational_Fishery

59 Sampling Design for Accuracy Assessment of Large-Area, Land ...

https://www.researchgate.net/publication/280296675_Sampling_Design_for_Accuracy_Assessment_of_Large-Area_Land-Cover_Maps

60 Sampling design for estimation of area and map accuracy - openMRV

https://openmrv.org/web/guest/w/modules/mrv/modules_3/sampling-design-for-estimation-of-area-and-map-accuracy

61 [PDF] Good Practices for Estimating Area and Assessing Accuracy of Land ...

https://nottingham-repository.worktribe.com/preview/728232/Olofsson_good%20practices.pdf