Instituto de Engenharia de Sistemas e Computadores de Coimbra

Institute of Systems Engineering and Computers

INESC - Coimbra

Ismael Jesus, Cidália Fonte, Jacinto Estima, Alberto Cardoso

Mathematical Proof of Thematic Accuracy Metrics used in Thematic Maps Accuracy Assessment

No. 1 2025

ISSN: 1645-2631

Instituto de Engenharia de Sistemas e Computadores de Coimbra

INESC - Coimbra

Rua Sílvio Lima, Polo II da UC, DEEC; 3030-790 Coimbra; Portugal

www.inescc.pt

Mathematical Proof of Thematic Accuracy Metrics used in Thematic Maps Accuracy Assessment

Ismael Jesus[1,2], Cidália C. Fonte[2], Jacinto Estima[1], Alberto Cardoso[1]

1. University of Coimbra, CISUC, Department of Informatics Engineering, Rua Sílvio Lima, 3030-290 Coimbra, Portugal
2. University of Coimbra, INESC Coimbra, Department of Mathematics, Largo D. Dinis, 3000 – 143 Coimbra, Portugal

Abstract

This technical report addresses the challenge of accurately assessing the quality of thematic maps, such as Land Use and/or Land Cover (LULC) maps, with focus on thematic accuracy. It presents the mathematical foundations underlying commonly used accuracy metrics, aiming to clarify how these metrics are derived and interpreted. The document is structured into two main sections: it begins with an overview of statistical estimators relevant to accuracy assessment, emphasizing simple and stratified random sampling techniques; then it defines key accuracy metrics, explains their relationship to the estimators, and it provides detailed derivations of the formulas for different sampling scenarios. This comprehensive approach supports a deeper understanding of the implications and reliability of thematic accuracy evaluations in geospatial analysis.

Keywords: Thematic accuracy assessment, confusion matrices, LULC maps, stratified random sampling, simple random sampling.

1 Introduction

Creating a map that accurately represents the real world is a significant challenge. The accuracy assessment of a thematic map - such as, a Land Use and/or Land Cover (LULC) map - entails several specific considerations, among which thematic accuracy is a key component. The main objective is to evaluate how closely the map reflects reality. Only with this understanding can we appropriately assess the risks involved in making decisions based on these maps. Thematic accuracy assessment involves comparing the thematic class(es) assigned to a geographic area on the map with the actual class(es) that should be assigned to those locations.

This technical report provides a comprehensive explanation of how the formulas for accuracy metrics are derived. Its goal is to help readers understand the mathematical foundations of the accuracy metrics commonly found in the literature, such as those presented by Olofsson et al. (2014) and available in appendix A. The document is organized into three main sections: Section 2 presents a brief overview of the key statistical estimators used to derive accuracy metrics, with a focus on simple random sampling and stratified random sampling designs; Section 3 defines the thematic accuracy metrics explaining how they relate to the estimator definitions, and offers detailed derivations of the formulas for each random sampling approach.

2 A Brief Overview of General Statistical Estimators

When it is necessary to infer the values a set of variables take regarding a population of individuals, entities, or elements in the statistical sense, it may be impossible to assess their exact value for all elements of the population. This may be due to several aspects, including the real feasibility of such task, or the cost and time associated with analyzing all the population units. In such cases sampling approaches may be used, which enable to estimate the statistical value we seek in relation to the population under study, based on the information collected for a reduced number of individuals selected from the population. The following sub-sections present a set of estimators that will be used in the demonstrations in section 3, grouped by the type of sampling employed: simple random sampling or stratified random sampling (e.g., Cochran, 1977).

2.1 Simple random sampling

Consider a population of size N, denoted by $U = \{u_1, u_2, \dots, u_N\}$. Assume that each unit in the population, u_L, where $L = 1,2,\dots,N$, is described by the variables Y and X, expressed as $Y(u_L) = Y_L$ and $X(u_L) = X_L$, respectively.

When a random sample of size n is drawn, denoted by $s = \{s_1, s_2, \dots, s_n\}$, where s_l is the l-th sample unit, the values of the variables Y and X for the l-th element in the sample are referred to as $Y(s_l) = y_l$ and $X(s_l) = x_l$, respectively. The sample means of the variables Y and X are denoted by $\bar{y}$ and $\bar{x}$, respectively, and are computed using equations (1) and (2). The samples variances of Y and X, represented by s_y^2 and s_x^2, respectively, are computed using equations (3) and (4), while the sample covariance between X and Y, denoted by s_{xy}, is obtained using equation (5). For more details, see Thompson (2012) and Cochran (1977).

$$\bar{y} = \frac{\sum_{l=1}^{n} y_l}{n} \qquad (1)$$

$$\bar{x} = \frac{\sum_{l=1}^{n} x_l}{n} \qquad (2)$$

$$s_y^2 = \sum_{l=1}^{n} \frac{(y_l - \bar{y})^2}{n-1} \qquad (3)$$

$$s_x^2 = \sum_{l=1}^{n} \frac{(x_l - \bar{x})^2}{n-1} \qquad (4)$$

$$s_{xy} = \sum_{l=1}^{n} \frac{(y_l - \bar{y})(x_l - \bar{x})}{n-1} \qquad (5)$$

When Y and X are dichotomous variables, that is, they can only take one of two values, such as 1 or 0, s_y^2 and s_x^2 can be simplified by the equations (6) and (7), respectively. Additionally, when Y implies X, i.e., $P(X = 1|Y = 1) = 1$, s_{xy} can be written by equation (8).

$$s_y^2 = \frac{n}{n-1}\bar{y}(1-\bar{y}) \qquad (6)$$

$$s_x^2 = \frac{n}{n-1}\bar{x}(1-\bar{x}) \qquad (7)$$

$$s_{xy} = \frac{n}{n-1}\bar{y}(1-\bar{x}) \qquad (8)$$

Table 1 summarizes the computation of the true values for the population total, mean, and ratio, along with their corresponding estimators based on samples obtained through simple random sampling. Note that the true value is calculated using all units, while the corresponding estimators and their variances are computed from the collected sample.

Total of *Y* for Population *U*	**True Value**	$$T_Y = \sum_{L=1}^{N} Y_L \quad (9)$$
	Estimator	$$\widehat{T_Y} = N\bar{y} \quad (10)$$
	Estimated Variance	$$\widehat{var}(\widehat{T_Y}) = \left(1 - \frac{n}{N}\right)\frac{N^2}{n} s_y^2 \quad (11)$$
Mean of *Y* for Population *U*	**True Value**	$$\bar{Y} = \frac{\sum_{L=1}^{N} Y_L}{N} \quad (12)$$
	Estimator	$$\hat{\bar{Y}} = \bar{y} \quad (13)$$
	Estimated Variance	$$\widehat{var}(\hat{\bar{Y}}) = \left(1 - \frac{n}{N}\right)\frac{1}{n} s_y^2 \quad (14)$$
Ratio between *Y* and *X* for Population *U*	**True Value**	$$R_{YX} = \frac{\sum_{L=1}^{N} Y_L}{\sum_{L=1}^{N} X_L} \quad (15)$$
	Estimator	$$\widehat{R_{YX}} = \frac{\bar{y}}{\bar{x}} \quad (16)$$
	Estimated Variance	$$\widehat{var}(\widehat{R_{YX}}) = \left(1 - \frac{n}{N}\right)\frac{1}{n\,\bar{x}^2}[s_y^2 + \widehat{R_{YX}}^2 s_x^2 - 2\widehat{R_{YX}} s_{xy}] \quad (17)$$

Table 1. Estimators and their estimated variances using simple random sampling.

2.2 Stratified random sampling

Consider a population of size N, denoted by $U = \{u_1, u_2, \dots, u_N\}$. Assume this population is partitioned into h disjointed strata, denoted by $S = \{S_1, S_2, \dots, S_h\}$, where the k-th stratum, S_k, has the size N_k^S. This set forms a partition of U, such that $U = \bigcup_{k=1}^{h} S_k$ and $S_p \cap S_q = \emptyset$, for all $p, q = 1,2, \dots, h$ with $p \neq q$. Assume each unit within the stratum S_k, with $k = 1,2, \dots, h$, denoted by u_{Lk}, where $L = 1,2, \dots, N_k^S$, can be described by the variables Y and X. These are denoted as $Y(u_{Lk}) = Y_{Lk}$ and $X(u_{Lk}) = X_{Lk}$, respectively.

When collecting a sample s of size n through stratified random sampling, let n_k represent the number of units of the sample in the stratum S_k. Specifically, let $s_k = s \cap S_k$ denote the sample from stratum S_k, where $s_k = \{s_{1k}, s_{2k}, \dots, s_{n_k k}\}$. For the l-th element of the stratum S_k, denoted as s_{lk}, we define the values of the variables Y and X as $Y(s_{lk}) = y_{lk}$ and $X(s_{lk}) = x_{lk}$, respectively.

Considering the stratum S_k, the sample means of the variables Y and X of kth-stratum are represented by $\bar{y}_k$ and $\bar{x}_k$, respectively, and are computed using equations (18) and (19). The sample variances of Y and X of kth-stratum, represented by s_{yk}^2 and s_{xk}^2, respectively, are computed using equations (20) and (21), while the sample covariance between X and Y of kth-stratum, denoted by s_{xyk}, is obtained using equation (22). For more details, see Thompson (2012) and Cochran (1977).

$$\bar{y}_k = \frac{\sum_{l=1}^{n_k} y_{lk}}{n_k} \quad (18)$$

$$\bar{x}_k = \frac{\sum_{l=1}^{n_k} x_{lk}}{n_k} \quad (19)$$

$$s_{yk}^2 = \sum_{l=1}^{n_k} \frac{(y_{lk} - \bar{y}_k)^2}{n_k - 1} \quad (20)$$

$$s_{xk}^2 = \sum_{l=1}^{n_k} \frac{(x_{lk} - \bar{x}_k)^2}{n_k - 1} \quad (21)$$

$$s_{xyk} = \sum_{l=1}^{n_k} \frac{(x_{lk} - \bar{x}_k)(y_{lk} - \bar{y}_k)}{n_k - 1} \quad (22)$$

When Y and X are dichotomous variables, s_{yk}^2 and s_{xk}^2, can be simplified by equations (23) and (24), respectively. Additionally, when Y implies X, i.e., $P(X = 1|Y = 1) = 1$, s_{xyk} can be written by equation (25).

$$s_{yk}^2 = \frac{n_k}{n_k - 1}\bar{y}_k(1 - \bar{y}_k) \quad (23)$$

$$s_{xk}^2 = \frac{n_k}{n_k - 1}\bar{x}_k(1 - \bar{x}_k) \quad (24)$$

$$s_{xyk} = \frac{n_k}{n_k - 1}\bar{y}_k(1 - \bar{x}_k) \quad (25)$$

Table 2 presents the formulas to compute the per stratum true total and mean, along with the corresponding estimators using a stratified random sample, as well as the true values for the population total, mean, and ratio, along with their corresponding estimators, based on a sample obtained through stratified random sampling. Note that the true value is calculated using all units, while the corresponding estimators and their variances are computed from the collected sample statistics.

Total of *Y* for Stratum S_k	**True Value**	$T_{Yk} = \sum_{L=1}^{N_k^S} Y_{Lk}$ *(26)*
	Estimator	$\widehat{T_{Yk}} = N_k^S \bar{y}_k$ *(27)*
	Estimated Variance	$\widehat{var}(\widehat{T_{Yk}}) = \left(1 - \frac{n_k}{N_k^S}\right)\frac{(N_k^S)^2}{n_k} s_{yk}^2$ *(28)*
Total of *Y* for Population *U*	**True Value**	$T_Y = \sum_{k=1}^{h}\sum_{L=1}^{N_k^S} Y_{Lk} = \sum_{k=1}^{h} T_{Yk}$ *(29)*
	Estimator	$\widehat{T_Y} = \sum_{k=1}^{h} \widehat{T_{Yk}}$ *(30)*

	Estimated Variance	$$\widehat{var}(\widehat{T_Y}) = \sum_{k=1}^{h} \widehat{var}(\widehat{T_{Yk}}) \quad (31)$$
Mean of Y for Stratum S_k	**True Value**	$$\overline{Y_k} = \frac{\sum_{L=1}^{N_k^S} Y_{Lk}}{N_k^S} \quad (32)$$
	Estimator	$$\widehat{\overline{Y_k}} = \bar{y}_k \quad (33)$$
	Estimated Variance	$$\widehat{var}\left(\widehat{\overline{Y_k}}\right) = \left(1 - \frac{n_k}{N_k^S}\right) \frac{1}{n_k} s_{yk}^2 \quad (34)$$
Mean of Y for Population U	**True Value**	$$\overline{Y} = \frac{\sum_{k=1}^{h} \sum_{L=1}^{N_k^S} Y_{Lk}}{N} = \sum_{k=1}^{h} \frac{N_k^S}{N} \overline{Y_k} \quad (35)$$
	Estimator	$$\widehat{\overline{Y}} = \sum_{k=1}^{h} \frac{N_k^S}{\mathrm{N}} \widehat{\overline{Y_k}} \quad (36)$$
	Estimated Variance	$$\widehat{var}\left(\widehat{\overline{Y}}\right) = \sum_{k=1}^{h} \left(\frac{N_k^S}{\mathrm{N}}\right)^2 \widehat{var}\left(\widehat{\overline{Y_k}}\right) \quad (37)$$
Ratio between Y and X for Population U	**True Value**	$$R_{YX} = \frac{\sum_{k=1}^{h} \sum_{L=1}^{N_k^S} Y_{Lk}}{\sum_{k=1}^{h} \sum_{L=1}^{N_k^S} X_{Lk}} \quad (38)$$
	Estimator	$$\widehat{R_{YX}} = \frac{\sum_{k=1}^{h} N_k^S \bar{y}_k}{\sum_{k=1}^{h} N_k^S \bar{x}_k} \quad (39)$$
	Estimated Variance	$$\widehat{var}(\widehat{R_{YX}}) = \frac{1}{(\sum_{k=1}^{h} N_k^S \,\bar{x}_k)^2} \sum_{k=1}^{h} \left(1 - \frac{n_k}{N_k^S}\right) * * \frac{(N_k^S)^2}{n_k} \left[s_{yk}^2 + \widehat{R_{YX}}^2 s_{xk}^2 - 2\widehat{R_{YX}} s_{xyk}\right] \quad (40)$$

Table 2. True values, estimators and their estimated variances of the variables total value, mean and ratio, per stratum and for all the population when using stratified random sampling.

3 Estimators for Thematic Accuracy Metrics

Estimators for thematic accuracy metrics can be based on the formulas presented in the previous section, using the appropriate definition of variables.

In the following subsections, we demonstrate how the mostly used parameters to evaluate thematic accuracy of a map are obtained. These are: the true number of units (e.g., pixels) in a class, the user's accuracy of a class, the producer's accuracy of a class, and the map overall accuracy. To compute such values, we use the definitions of the estimators and the relationship between the accuracy metrics and the statistic estimators presented in previous section.

After completing the sample collection and the response design phase (see e.g., Stehman and Foody, 2019), each selected sampling unit is assigned to only one map class and one reference class[1]. This information may be organized in a confusion matrix, as illustrated in Figure 1, where the value n_{ij} in cell (i, j) is the number of sampling units classified as class i on the map and class j in the reference data. The marginal totals are defined as $n_{i+} = \sum_{j=1}^{q} n_{ij}$ and $n_{+j} = \sum_{i=1}^{q} n_{ij}$, representing the total number of sampling units classified as class i on the map and class j in the reference data, respectively. It is important to note that $\sum_{i=1}^{q} n_{i+} = \sum_{j=1}^{q} n_{+j} = \sum_{i=1}^{q} \sum_{j=1}^{q} n_{ij} = n$.

Map \ Reference	**Class 1**	**Class 2**	...	**Class j**	...	**Class q**	**Total**
Class 1	n_{11}	n_{12}	...	n_{1j}	...	n_{1q}	$\boldsymbol{n_{1+}}$
Class 2	n_{21}	n_{22}	...	n_{2j}	...	n_{2q}	$\boldsymbol{n_{2+}}$
...	...	...	...	...	...	...	...
Class i	n_{i1}	n_{i2}	...	n_{ij}	...	n_{iq}	$\boldsymbol{n_{i+}}$
...	...	...	...	...	...	...	...
Class q	n_{q1}	n_{q2}	...	n_{qj}	...	n_{qq}	$\boldsymbol{n_{q+}}$
Total	$\boldsymbol{n_{+1}}$	$\boldsymbol{n_{+2}}$	...	$\boldsymbol{n_{+j}}$	...	$\boldsymbol{n_{+q}}$	$\boldsymbol{n}$

Figure 1. Confusion matrix, where n_{ij} in each cell (i,j) represents the number of sampling units classified as class i on the map and class j in the reference data. The marginal totals, n_{i+} and n_{+j}, correspond to the total number of sampling units classified as class i on the map and class j in the reference data, respectively.

When using stratified random sampling, considering the classes as strata, in the formulas presented in Section 2.2, we have that:

- h corresponds to the number of strata, which, in this case, is also the number of classes, denoted as q;
- N_k^S corresponds the total number of units in the kth-stratum, which, in this case, is the total number of spatial units (e.g., pixels) in the kth-class on the map, is denoted as N_k^M;
- n_k represents the number of selected sampling units of the kth-stratum, which corresponds to the number of selected sampling units in the kth-class of the map, denoted as n_{k+}.

[1] In fact, in some cases more than one reference class may be associated with each sample unit. However, in this document it is considered that in such cases one class from the possible selected classes will be chosen as reference class, corresponding, for example, to the class occupying the largest area of the spatial unit or the class considered as more representative of the spatial unit according to the map specifications. For more details on this topic see, e.g., Fonte et al. (2020) and Stehman and Foody (2019).

3.1 Estimation of the True Number of Units in a Class

The true number of units in a class j, N_j^R, is the number of units that belong to class j in the real world and it can be computed using formula (41):

$$N_j^R = \sum_{u \in U} Y(u)\,, \qquad (41)$$

where, Y is a dichotomous variable that takes the value 1 when the unit u belongs to class j in the reference and 0 otherwise.

This measure can be estimated by a total estimator, (10) or (30), depending on the sampling design used, and using as definition of Y:

$$Y(u) = \begin{cases} 1 & , if \text{ class } j \; is\ assigned\ to\ u\ in\ the\ reference \\ 0 & , otherwise \end{cases} \qquad (42)$$

Note that the sum of the variable Y represents the number of units that belong to class j in the reference. For example, for a set A of spatial units within the region of interest U, $\sum_{u \in \mathrm{A}} Y(u)$ gives us the number of units in set A assigned to class j in the reference.

3.1.1 Simple random sampling

Given the above definition of Y, equation (42), the sample mean of the variable Y can be computed using equation (1). That is:

$$\begin{aligned} \bar{y} &= \frac{\sum_{l=1}^{n} y_l}{n} \\ &= \frac{n_{+j}}{n} \end{aligned}$$

Using equation (10), the estimated real number of spatial units in class j, $\widehat{N_j^R}$, is computed by:

$$\begin{aligned} \widehat{N_j^R} &= N\bar{y} \\ &= N\frac{n_{+j}}{n} \end{aligned}$$

As Y is a dichotomous variable, equation (42), the sample variance of the variable Y can be computed using equation (6). Therefore, we have:

$$\begin{aligned} s_y^2 &= \frac{n}{n-1}\bar{y}(1-\bar{y}) \\ &= \frac{n}{n-1}\frac{n_{+j}}{n}\left(1-\frac{n_{+j}}{n}\right) \\ &= \frac{n_{+j}}{n-1}\left(1-\frac{n_{+j}}{n}\right) \end{aligned}$$

Using equation (11), the estimated variance of the estimated real number of units in class j can be computed using:

$$\begin{aligned}\widehat{var}\left(\widehat{N_j^R}\right) &= \left(1-\frac{n}{N}\right)\frac{N^2}{n}s_y^2\\ &= \left(1-\frac{n}{N}\right)\frac{N^2}{n}\frac{n_{+j}}{n-1}\left(1-\frac{n_{+j}}{n}\right)\\ &= \left(1-\frac{n}{N}\right)\frac{1}{n-1}N\frac{n_{+j}}{n}\left(N-N\frac{n_{+j}}{n}\right)\\ &= \left(1-\frac{n}{N}\right)\frac{\widehat{N_j^R}\left(N-\widehat{N_j^R}\right)}{n-1}\end{aligned}$$

Assuming $n \ll N$, that is, the population size is much larger than the size of the sample, then $\left(1-\frac{n}{N}\right) \approx 1$ and:

$$\widehat{var}\left(\widehat{N_j^R}\right) = \frac{\widehat{N_j^R}\left(N-\widehat{N_j^R}\right)}{n-1}$$

3.1.2 *Stratified random sampling based on map classes*

Given the definition of Y, equation (42), the sample mean of variable Y for the kth-stratum (i.e., kth-class) can be computed with equation (18). Therefore, we have:

$$\begin{aligned}\bar{y}_k &= \frac{\sum_{l=1}^{n_{k+}} y_{lk}}{n_{k+}}\\ &= \frac{n_{kj}}{n_{k+}}\end{aligned}$$

The estimated real number of spatial units in class j is computed using equations (27) and (30). Therefore, we have:

$$\begin{aligned}\widehat{N_j^R} &= \sum_{k=1}^{q} N_k^M \bar{y}_k\\ &= \sum_{k=1}^{q} N_k^M \frac{n_{kj}}{n_{k+}}\end{aligned} \qquad (43)$$

As Y is a dichotomous variable, equation (42), the sample variance of variable Y for the kth-stratum (i.e., kth-class) can be computed using equation (23). That is:

$$\begin{aligned}s_{yk}^2 &= \frac{n_{k+}}{n_{k+}-1}\bar{y}_k(1-\bar{y}_k)\\ &= \frac{n_{k+}}{n_{k+}-1}\frac{n_{kj}}{n_{k+}}\left(1-\frac{n_{kj}}{n_{k+}}\right)\\ &= \frac{n_{kj}}{n_{k+}-1}\left(1-\frac{n_{kj}}{n_{k+}}\right)\end{aligned}$$

The estimated variance of the estimated real number of units in class j can be computed using equations (28) to estimate the per map class variance, and equation (31) for the population variance. Then:

$$\begin{aligned}\widehat{var}\left(\widehat{N_j^R}\right) &= \sum_{k=1}^{q}\left(1-\frac{n_{k+}}{N_k^M}\right)\frac{(N_k^M)^2}{n_{k+}}s_{yk}^2\\ &= \sum_{k=1}^{q}\left(1-\frac{n_{k+}}{N_k^M}\right)\frac{(N_k^M)^2}{n_{k+}-1}\frac{n_{kj}}{n_{k+}}\left(1-\frac{n_{kj}}{n_{k+}}\right)\end{aligned} \qquad (44)$$

Assuming that $n_k \ll N_k^M$, that is, the stratum size is much larger than the size of the sample in that stratum, we have that $\left(1-\frac{n_k}{N_k^M}\right) \approx 1$, and therefore:

$$\widehat{var}(\widehat{N_j^R}) = \sum_{k=1}^{q} \frac{(N_k^M)^2}{n_{k+}-1} \frac{n_{kj}}{n_{k+}}\left(1 - \frac{n_{kj}}{n_{k+}}\right)$$

Whenever the sample is drawn considering the number of sample units per stratum (which in this case are the classes) proportional to the strata (classes') area, i.e., assuming the condition:

$$\frac{N}{n} = \frac{N_K^M}{n_{k+}}$$

the estimated real number of units in class j can be written as:

$$\begin{aligned} \widehat{N_j^R} &= \sum_{k=1}^{q} N_k^M \frac{n_{kj}}{n_{k+}} \\ &= \sum_{k=1}^{q} \frac{N}{n} n_{kj} \\ &= \frac{N}{n} \sum_{k=1}^{q} n_{kj} \\ &= \frac{N}{n} n_{+j} \end{aligned} \tag{45}$$

and the estimated variance of the estimated real number of units in class j; can be computed with:

$$\begin{aligned} \widehat{var}(\widehat{N_j^R}) &= \sum_{k=1}^{q} \left(1 - \frac{n_{k+}}{N_k^M}\right) \frac{(N_k^M)^2}{n_{k+}} \frac{n_{kj}}{n_{k+}-1}\left(1 - \frac{n_{kj}}{n_{k+}}\right) \\ &= \sum_{k=1}^{q} \left(1 - \frac{n}{N}\right) n_{k+} \left(\frac{N}{n}\right)^2 \frac{n_{kj}}{n_{k+}-1}\left(1 - \frac{n_{kj}}{n_{k+}}\right) \\ &= \left(1 - \frac{n}{N}\right)\left(\frac{N}{n}\right)^2 \sum_{k=1}^{q} \frac{n_{k+}}{n_{k+}-1} n_{kj} \left(1 - \frac{n_{kj}}{n_{k+}}\right) \end{aligned}$$

Assuming $n \ll N$, then $\left(1 - \frac{n}{N}\right) \approx 1$ and:

$$\widehat{var}(\widehat{N_j^R}) = \left(\frac{N}{n}\right)^2 \sum_{k=1}^{q} \frac{n_{k+}}{n_{k+}-1} n_{kj} \left(1 - \frac{n_{kj}}{n_{k+}}\right)$$

3.2 Estimation of the User's Accuracy of a Class

User's accuracy of class i, UA_i, is the proportion between the number of units correctly classified in class i and the total number of units that were assigned to class i on the map. It can be computed using equation (46):

$$UA_i = \frac{\sum_{u \in U} Y(u)}{\sum_{u \in U} X(u)} \tag{46}$$

where: Y is defined as a dichotomous variable that takes the value 1 when the unit u belongs to class i both in the reference and on the map, and 0 otherwise; and X is defined as a dichotomous variable that takes the value 1 when the unit u belongs to class i on the map, and 0 otherwise. That is:

$$Y(u) = \begin{cases} 1 & , if\ class\ i\ assigned\ is\ assigned\ to\ u\ in\ the\ reference\ and\ the\ map \\ 0 & , otherwise \end{cases} \tag{47}$$

$$X(u) = \begin{cases} 1 & , if\ class\ i\ is\ assigned\ to\ u\ in\ the\ map \\ 0 & , otherwise \end{cases} \quad (48)$$

To estimate the value of this measure based on a sample, a ratio estimator can be used. Therefore, the estimators presented in equations (16) or (39) may be used, respectively, when simple random samples or stratified random samples are considered.

Note that according to the above definitions of Y and X, equations (47) and (48), respectively, the sum of the variable Y represents the number of units that belong to class i both on the map and in the reference. Similarly, the sum of the variable X represents the number of units that belong to class i on the map.

3.2.1 Simple random sampling

When considering a simple random sample, given to the definition of Y, equation (47), and using equation (1), the sample mean of the variable Y is given by:

$$\begin{aligned} \bar{y} &= \frac{\sum_{l=1}^{n} y_l}{n} \\ &= \frac{n_{ii}}{n} \end{aligned}$$

Considering the above the definition of X, equation (48), and using equation (2), the sample mean of the variable X is given by:

$$\begin{aligned} \bar{x} &= \frac{\sum_{l=1}^{n} x_l}{n} \\ &= \frac{n_{i+}}{n} \end{aligned}$$

Using equation (16), the estimated user's accuracy of class i is computed by:

$$\begin{aligned} \widehat{UA}_i &= \frac{\bar{y}}{\bar{x}} \\ &= \frac{\frac{n_{ii}}{n}}{\frac{n_{i+}}{n}} \\ &= \frac{n_{ii}}{n_{i+}} \end{aligned}$$

As Y is a dichotomous variable, equation (47), using equation (6), the sample variance of the variable Y is given by:

$$\begin{aligned} s_y^2 &= \frac{n}{n-1}\bar{y}(1-\bar{y}) \\ &= \frac{n}{n-1}\frac{n_{ii}}{n}\left(1-\frac{n_{ii}}{n}\right) \\ &= \frac{n_{ii}}{n-1}\left(1-\frac{n_{ii}}{n}\right) \end{aligned}$$

As X is a dichotomous variable, equation (48), using equation (7), the sample variance of the variable X is given by:

$$
\begin{aligned}
s_x^2 &= \frac{n}{n-1}\bar{x}(1-\bar{x}) \\
&= \frac{n}{n-1}\frac{n_{i+}}{n}\left(1-\frac{n_{i+}}{n}\right) \\
&= \frac{n_{i+}}{n-1}\left(1-\frac{n_{i+}}{n}\right)
\end{aligned}
$$

As Y and X are dichotomous variables, equations (47) and (48), respectively, and $Y = 1$ implies $X = 1$, using equation (8), the sample covariance between the variables X and Y is given by:

$$
\begin{aligned}
s_{xy} &= \frac{n}{n-1}\bar{y}(1-\bar{x}) \\
&= \frac{n}{n-1}\frac{n_{ii}}{n}\left(1-\frac{n_{i+}}{n}\right) \\
&= \frac{n_{ii}}{n-1}\left(1-\frac{n_{i+}}{n}\right)
\end{aligned}
$$

Using equation (17), the estimated variance of the estimated user's accuracy of class i is computed by:

$$
\begin{aligned}
\widehat{var}\left(\widehat{UA}_i\right) &= \left(1-\frac{n}{N}\right)\frac{1}{n\,\bar{x}^2}\left[s_y^2 + \widehat{UA}_i^{\,2} s_x^2 - 2\widehat{UA}_i s_{xy}\right] \\
&= \left(1-\frac{n}{N}\right)\frac{1}{n\left(\frac{n_{i+}}{n}\right)^2}\left[\frac{n_{ii}}{n-1}\left(1-\frac{n_{ii}}{n}\right) + \widehat{UA}_i^{\,2}\frac{n_{i+}}{n-1}\left(1-\frac{n_{i+}}{n}\right) - 2\widehat{UA}_i\frac{n_{ii}}{n-1}\left(1-\frac{n_{i+}}{n}\right)\right] \\
&= \left(1-\frac{n}{N}\right)\frac{n}{n_{i+}^2}\frac{1}{n-1}\left[n_{ii}\left(1-\frac{n_{ii}}{n}\right) + \widehat{UA}_i^{\,2} n_{i+}\left(1-\frac{n_{i+}}{n}\right) - 2\widehat{UA}_i n_{ii}\left(1-\frac{n_{i+}}{n}\right)\right] \\
&= \left(1-\frac{n}{N}\right)\frac{n}{n_{i+}}\frac{1}{n-1}\left[\frac{n_{ii}}{n_{i+}}\left(1-\frac{n_{ii}}{n}\right) + \widehat{UA}_i^{\,2}\frac{n_{i+}}{n_{i+}}\left(1-\frac{n_{i+}}{n}\right) - 2\widehat{UA}_i\frac{n_{ii}}{n_{i+}}\left(1-\frac{n_{i+}}{n}\right)\right] \\
&= \left(1-\frac{n}{N}\right)\frac{n}{n-1}\frac{1}{n_{i+}}\left[\widehat{UA}_i\left(1-\frac{n_{ii}}{n}\right) + \widehat{UA}_i^{\,2}\left(1-\frac{n_{i+}}{n}\right) - 2\widehat{UA}_i^2\left(1-\frac{n_{i+}}{n}\right)\right] \\
&= \left(1-\frac{n}{N}\right)\frac{n}{n-1}\frac{1}{n_{i+}}\left[\widehat{UA}_i\left(1-\frac{n_{ii}}{n}\right) - \widehat{UA}_i^{\,2}\left(1-\frac{n_{i+}}{n}\right)\right] \\
&= \left(1-\frac{n}{N}\right)\frac{n}{n-1}\frac{\widehat{UA}_i}{n_{i+}}\left[\left(1-\frac{n_{ii}}{n}\right) - \widehat{UA}_i\left(1-\frac{n_{i+}}{n}\right)\right] \\
&= \left(1-\frac{n}{N}\right)\frac{n}{n-1}\frac{\widehat{UA}_i}{n_{i+}}\left[1-\frac{n_{ii}}{n} - \widehat{UA}_i + \widehat{UA}_i\frac{n_{i+}}{n}\right] \\
&= \left(1-\frac{n}{N}\right)\frac{n}{n-1}\frac{\widehat{UA}_i}{n_{i+}}\left[1-\frac{n_{ii}}{n} - \widehat{UA}_i + \frac{n_{ii}}{n_{i+}}\frac{n_{i+}}{n}\right] \\
&= \left(1-\frac{n}{N}\right)\frac{n}{n-1}\frac{\widehat{UA}_i}{n_{i+}}\left(1-\frac{n_{ii}}{n} - \widehat{UA}_i + \frac{n_{ii}}{n}\right) \\
&= \left(1-\frac{n}{N}\right)\frac{n}{n-1}\frac{\widehat{UA}_i\left(1-\widehat{UA}_i\right)}{n_{i+}}
\end{aligned}
$$

Assuming $n \ll N$, i.e., the population size is much larger than the size of the sample, $\left(1-\frac{n}{N}\right) \approx 1$ and:

$$
\widehat{var}\left(\widehat{UA}_i\right) = \frac{n}{n-1}\frac{\widehat{UA}_i\left(1-\widehat{UA}_i\right)}{n_{i+}}
$$

Also, if the sample is large, the ratio $\frac{n}{n-1}$ is approximately equal to 1, and

$$
\widehat{var}\left(\widehat{UA}_i\right) = \frac{\widehat{UA}_i\left(1-\widehat{UA}_i\right)}{n_{i+}}
$$

3.2.2 *Stratified random sampling based on map classes*

Assuming the above definition of Y, equation (47), and using equation (18), the sample mean of variable Y for the kth-stratum (i.e., a set of spatial units corresponding to the kth-class on the map) is given by:

$$\bar{y}_k = \frac{\sum_{i=1}^{n_{k+}} y_{ik}}{n_{k+}} = \begin{cases} \dfrac{0}{n_{k+}} & , if \text{ class } k \neq i \\ \dfrac{n_{ii}}{n_{i+}} & , if \text{ class } k = i \end{cases} = \begin{cases} 0 & , if \text{ class } k \neq i \\ \dfrac{n_{ii}}{n_{i+}} & , if \text{ class } k = i \end{cases}$$

Considering the above definition of X, equation (48), and using equation (19), the sample mean of the variable X for the kth-stratum (i.e., kth-class) is given by:

$$\bar{x}_k = \frac{\sum_{i=1}^{n_{k+}} x_{ik}}{n_{k+}} = \begin{cases} \dfrac{0}{n_{k+}} & , if \text{ class } k \neq i \\ \dfrac{n_{i+}}{n_{i+}} & , if \text{ class } k = i \end{cases} = \begin{cases} 0 & , if \text{ class } k \neq i \\ 1 & , if \text{ class } k = i \end{cases}$$

Using equation (39), the estimated user's accuracy of class i is computed by:

$$\widehat{UA}_i = \frac{\sum_{k=1}^{q} N_k^M \bar{y}_k}{\sum_{k=1}^{q} N_k^M \bar{x}_k} = \frac{N_1^M \bar{y}_1 + \cdots + N_{i-1}^M \bar{y}_{i-1} + N_i^M \bar{y}_i + N_{i+1}^M \bar{y}_{i+1} + \cdots + N_q^M \bar{y}_q}{N_1^M \bar{x}_q + \cdots + N_{i-1}^M \bar{x}_{i-1} + N_i^M \bar{x}_i + N_{i+1}^M \bar{x}_{i+1} + \cdots + N_q^M \bar{x}_q} = \frac{N_1^M 0 + \cdots + N_{i-1}^M 0 + N_i^M \frac{n_{ii}}{n_{i+}} + N_{i+1}^M 0 + \cdots + N_q^M 0}{N_1^M 0 + \cdots + N_{i-1}^M 0 + N_i^M 1 + N_{i+1}^M 0 + \cdots + N_q^M 0} = \frac{N_i^M \frac{n_{ii}}{n_{i+}}}{N_i^M} = \frac{n_{ii}}{n_{i+}}$$

As Y is a dichotomous variable, equation (47), using equation (23), the sample variance of variable Y for the kth-stratum (i.e., kth-class) is given by:

$$s_{yk}^2 = \frac{n_{k+}}{n_{k+} - 1} \bar{y}_k (1 - \bar{y}_k) = \begin{cases} \dfrac{n_{k+}}{n_{k+} - 1} 0(1 - 0) & , if \text{ class } k \neq i \\ \dfrac{n_{i+}}{n_{i+} - 1} \dfrac{n_{ii}}{n_{i+}} \left(1 - \dfrac{n_{ii}}{n_{i+}}\right) & , if \text{ class } k = i \end{cases} = \begin{cases} 0 & , if \text{ class } k \neq i \\ \dfrac{n_{ii}}{n_{i+} - 1} \left(1 - \dfrac{n_{ii}}{n_{i+}}\right) & , if \text{ class } k = i \end{cases}$$

As X is a dichotomous variable, equation (48), using equation (24), the sample variance of the variable X for the kth-stratum (i.e., kth-class) is given by:

$$
\begin{aligned}
s_{xk}^2 &= \frac{n_{k+}}{n_{k+}-1}\bar{x}_k(1-\bar{x}_k) \\
&= \begin{cases} \dfrac{n_{k+}}{n_{k+}-1}0(1-0) & ,if \text{ class } k \neq i \\ \dfrac{n_{k+}}{n_{k+}-1}1(1-1) & ,if \text{ class } k = i \end{cases} \\
&= 0
\end{aligned}
$$

As Y and X are dichotomous variables, equations (47) and (48), respectively, using equation (25), the sample covariance between the variables X and Y of kth-stratum (i.e., kth-class) is given by:

$$
\begin{aligned}
s_{xyk} &= \frac{n_{k+}}{n_{k+}-1}\bar{y}_k(1-\bar{x}_k) \\
&= \begin{cases} \dfrac{n_{k+}}{n_{k+}-1}0(1-0) & ,if \text{ class } k \neq i \\ \dfrac{n_{k+}}{n_{k+}-1}\dfrac{n_{ii}}{n_{i+}}(1-1) & ,if \text{ class } k = i \end{cases} \\
&= 0
\end{aligned}
$$

Using equation (40), the estimated variance of the estimated user's accuracy of class i is computed by:

$$
\begin{aligned}
(49) &= \frac{1}{(\sum_{k=1}^{h} N_k^M \bar{x}_k)^2}\sum_{k=1}^{q}\left(1-\frac{n_{k+}}{N_k^M}\right)\frac{(N_k^M)^2}{n_{k+}}\left[s_{yk}^2 + \widehat{UA}_i^{\,2} s_{xk}^2 - 2\widehat{UA}_i s_{xyk}\right] \\
&= \frac{1}{(N_i^M \bar{x}_i)^2}\sum_{k=1}^{q}\left(1-\frac{n_{k+}}{N_k^M}\right)\frac{(N_k^M)^2}{n_{k+}}\left[s_{yk}^2 + \widehat{UA}_i^{\,2} s_{xk}^2 - 2\widehat{UA}_i s_{xyk}\right] \\
&= \frac{1}{(N_i^M 1)^2}\sum_{k=1}^{q}\left(1-\frac{n_{k+}}{N_k^M}\right)\frac{(N_k^M)^2}{n_{k+}}\left[s_{yk}^2 + \widehat{UA}_i^{\,2} 0 - 2\widehat{UA}_i 0\right] \\
&= \frac{1}{(N_i^M)^2}\sum_{k=1}^{q}\left(1-\frac{n_{k+}}{N_k^S}\right)\frac{(N_k^M)^2}{n_{k+}} s_{yk}^2 \\
&= \frac{1}{(N_i^M)^2}(N_i^M)^2\left(1-\frac{n_{i+}}{N_i^M}\right)\frac{1}{n_{i+}}\frac{n_{ii}}{n_{i+}-1}\left(1-\frac{n_{ii}}{n_{i+}}\right) \\
&= \left(1-\frac{n_{i+}}{N_i^M}\right)\frac{1}{n_{i+}-1}\frac{n_{ii}}{n_{i+}}\left(1-\frac{n_{ii}}{n_{i+}}\right) \\
&= \left(1-\frac{n_{i+}}{N_i^M}\right)\frac{\widehat{UA}_i(1-\widehat{UA}_i)}{n_{i+}-1}
\end{aligned}
\qquad (49)
$$

Assuming $n_k \ll N_k^M$, i.e., the strata size is much larger than the size of the sample selected in that stratum, $\left(1-\frac{n_k}{N_k^M}\right) \approx 1$ and:

$$
\widehat{var}(\widehat{UA}_i) = \frac{\widehat{UA}_i(1-\widehat{UA}_i)}{n_{i+}-1}
$$

Whenever the sample size is proportional to the area of the map classes, i.e., assuming the condition:

$$
\frac{N}{n} = \frac{N_K^M}{n_{k+}}
$$

the estimated user's accuracy of class i is still computed with:

$$\widehat{UA}_i = \frac{n_{ii}}{n_{i+}}$$

But the estimated variance of the estimated user's accuracy of class i, when it cannot be assumed that $n_k \ll N_k^M$, is now computed with:

$$\begin{aligned} \widehat{var}(\widehat{UA}_i) &= \left(1 - \frac{n_{i+}}{N_i^M}\right) \frac{\widehat{UA}_i(1 - \widehat{UA}_i)}{n_{i+} - 1} \\ &= \left(1 - \frac{n}{N}\right) \frac{\widehat{UA}_i(1 - \widehat{UA}_i)}{n_{i+} - 1} \end{aligned}$$

Assuming $n \ll N$, then $\left(1 - \frac{n}{N}\right) \approx 1$ and:

$$\widehat{var}(\widehat{UA}_i) = \frac{\widehat{UA}_i(1 - \widehat{UA}_i)}{n_{i+} - 1}$$

3.3 Estimation of the Producer's Accuracy of a Class

Producer's accuracy of the class j, PA_j, is the proportion of units correctly classified as belonging to class j, relative to the total number of units that actually belongs to class j in the reference. This measure can be computed using equation (50):

$$PA_j = \frac{\sum_{u \in U} Y(u)}{\sum_{u \in U} X(u)}, \qquad (50)$$

where: Y is defined as a dichotomous variable that takes the value 1 when the unit u belongs to class j in both the reference and the map, and 0 otherwise; and X is defined as a dichotomous variable that takes the value 1 when the unit u belongs to the class j in the reference, and 0 otherwise. That is:

$$Y(u) = \begin{cases} 1 & , if \text{ class } j \text{ } is\ assigned\ to\ u\ in\ the\ reference\ and\ the\ map \\ 0 & , otherwise \end{cases} \qquad (51)$$

$$X(u) = \begin{cases} 1 & , if \text{ class } j \text{ } is\ assigned\ to\ u\ in\ the\ reference \\ 0 & , otherwise \end{cases} \qquad (52)$$

This measure can also be estimated by the ratio estimators (16) or (39), depending on the sampling design used, considering the above definitions of Y and X, equations (51) and (52), respectively. Note that in this case the sum of the variable Y still represents the number of units that belong to class j both on the map and in the reference. However, the sum of the variable X with the above definition, equation (52), represents the number of units that belong to class j in the reference. That is, for a set A of spatial units u, $\sum_{u \in \mathrm{A}} X(u)$ gives the number of units in set A assigned to class j in the reference.

3.3.1 Simple random sampling

In this case, given the definition of Y, equation (51), and using equation (1), the sample mean of the variable Y is given by:

$$\begin{aligned} \bar{y} &= \frac{\sum_{l=1}^{n} y_l}{n} \\ &= \frac{n_{jj}}{n} \end{aligned}$$

According to the definition of X, equation (52), and using equation (2), the sample mean of the variable X is given by:

$$\begin{aligned}\bar{x} &= \frac{\sum_{l=1}^{n} x_l}{n} \\ &= \frac{n_{+j}}{n}\end{aligned}$$

Using equation (16), the estimated producer's accuracy of class j is computed by:

$$\begin{aligned}\widehat{PA}_j &= \frac{\bar{y}}{\bar{x}} \\ &= \frac{\frac{n_{jj}}{n}}{\frac{n_{+j}}{n}} \\ &= \frac{n_{jj}}{n_{+j}}\end{aligned}$$

As Y is a dichotomous variable, equation (51), using equation (6), the sample variance of the variable Y is given by:

$$\begin{aligned}s_y^2 &= \frac{n}{n-1}\bar{y}(1-\bar{y}) \\ &= \frac{n}{n-1}\frac{n_{jj}}{n}\left(1-\frac{n_{jj}}{n}\right) \\ &= \frac{n_{jj}}{n-1}\left(1-\frac{n_{jj}}{n}\right)\end{aligned}$$

As X is also a dichotomous variable, equation (52), using equation (7), the sample variance of the variable X is given by:

$$\begin{aligned}s_x^2 &= \frac{n}{n-1}\bar{x}(1-\bar{x}) \\ &= \frac{n}{n-1}\frac{n_{+j}}{n}\left(1-\frac{n_{+j}}{n}\right) \\ &= \frac{n_{+j}}{n-1}\left(1-\frac{n_{+j}}{n}\right)\end{aligned}$$

Given that both Y and X are dichotomous variables, equations (51) and (52), respectively, using equation (8), the sample covariance between these variables X and Y is given by:

$$\begin{aligned}s_{xy} &= \frac{n}{n-1}\bar{y}(1-\bar{x}) \\ &= \frac{n}{n-1}\frac{n_{jj}}{n}\left(1-\frac{n_{+j}}{n}\right) \\ &= \frac{n_{jj}}{n-1}\left(1-\frac{n_{+j}}{n}\right)\end{aligned}$$

Using equation (17), the estimated variance of the estimated producer's accuracy of class j is computed by:

$$
\begin{aligned}
\widehat{var}\left(\widehat{PA}_j\right) &= \left(1-\frac{n}{N}\right)\frac{1}{n\,\bar{x}^2}\left[s_y^2+\widehat{PA}_j^{\,2}s_x^2-2\widehat{PA}_{ji}s_{xy}\right]\\
&= \left(1-\frac{n}{N}\right)\frac{1}{n\left(\frac{n_{+j}}{n}\right)^2}\left[\frac{n_{jj}}{n-1}\left(1-\frac{n_{jj}}{n}\right)+\widehat{PA}_j^{\,2}\frac{n_{+j}}{n-1}\left(1-\frac{n_{+j}}{n}\right)-2\widehat{PA}_j\frac{n_{jj}}{n-1}\left(1-\frac{n_{+j}}{n}\right)\right]\\
&= \left(1-\frac{n}{N}\right)\frac{n}{n_{+j}^2}\frac{1}{n-1}\left[n_{jj}\left(1-\frac{n_{jj}}{n}\right)+\widehat{PA}_j^{\,2}n_{+j}\left(1-\frac{n_{+j}}{n}\right)-2\widehat{PA}_j n_{jj}\left(1-\frac{n_{+j}}{n}\right)\right]\\
&= \left(1-\frac{n}{N}\right)\frac{n}{n_{+j}}\frac{1}{n-1}\left[\frac{n_{jj}}{n_{+j}}\left(1-\frac{n_{jj}}{n}\right)+\widehat{PA}_j^{\,2}\left(1-\frac{n_{+j}}{n}\right)-2\widehat{PA}_j\frac{n_{jj}}{n_{+j}}\left(1-\frac{n_{+j}}{n}\right)\right]\\
&= \left(1-\frac{n}{N}\right)\frac{n}{n-1}\frac{1}{n_{+j}}\left[\widehat{PA}_j\left(1-\frac{n_{jj}}{n}\right)+\widehat{PA}_j^{\,2}\left(1-\frac{n_{+j}}{n}\right)-2\widehat{PA}_j^2\left(1-\frac{n_{+j}}{n}\right)\right]\\
&= \left(1-\frac{n}{N}\right)\frac{n}{n-1}\frac{1}{n_{+j}}\left[\widehat{PA}_j\left(1-\frac{n_{jj}}{n}\right)-\widehat{PA}_j^{\,2}\left(1-\frac{n_{+j}}{n}\right)\right]\\
&= \left(1-\frac{n}{N}\right)\frac{n}{n-1}\frac{\widehat{PA}_j}{n_{+j}}\left[1-\frac{n_{jj}}{n}-\widehat{PA}_j+\frac{n_{+j}}{n}\widehat{PA}_j\right]\\
&= \left(1-\frac{n}{N}\right)\frac{n}{n-1}\frac{\widehat{PA}_j}{n_{+j}}\left[1-\frac{n_{jj}}{n}-\widehat{PA}_j+\frac{n_{+j}}{n}\frac{n_{jj}}{n_{+j}}\right]\\
&= \left(1-\frac{n}{N}\right)\frac{n}{n-1}\frac{\widehat{PA}_j}{n_{+j}}\left[1-\frac{n_{jj}}{n}-\widehat{PA}_j+\frac{n_{jj}}{n}\right]\\
&= \left(1-\frac{n}{N}\right)\frac{n}{n-1}\frac{\widehat{PA}_j\left(1-\widehat{PA}_j\right)}{n_{+j}}
\end{aligned}
$$

Assuming $n \ll N$, i.e., the population size is much larger than the size of the sample, $\left(1-\frac{n}{N}\right)\approx 1$ and:

$$
\widehat{var}\left(\widehat{PA}_j\right)=\frac{n}{n-1}\frac{\widehat{PA}_j\left(1-\widehat{PA}_j\right)}{n_{+j}}
$$

Also, if the sample is large, the ratio $\frac{n}{n-1}$ is approximately equal to 1, and

$$
\widehat{var}\left(\widehat{PA}_j\right)=\frac{\widehat{PA}_j\left(1-\widehat{PA}_j\right)}{n_{+j}}
$$

3.3.2 Stratified random sampling based on map classes

Given the above definition of Y, equation (51), and using equation (18), the sample mean of variable Y for the kth-stratum (i.e., kth-class) is given by:

$$
\begin{aligned}
\bar{y}_k &= \frac{\sum_{l=1}^{n_{k+}} y_{lk}}{n_{k+}}\\
&= \begin{cases}\dfrac{0}{n_{k+}} & ,if \text{ class } k \neq j\\[2ex] \dfrac{n_{ii}}{n_{j+}} & ,if \text{ class } k = j\end{cases}\\
&= \begin{cases}0 & ,if \text{ class } k \neq j\\ \dfrac{n_{jj}}{n_{j+}} & ,if \text{ class } k = j\end{cases}
\end{aligned}
$$

Considering the definition of X, equation (52), and using equation (19), the sample mean of the variable X for the kth-stratum (i.e., kth-class) is given by:

$$\bar{x}_k = \frac{\sum_{l=1}^{n_{k+}} x_{lk}}{n_{k+}} = \frac{n_{kj}}{n_{k+}}$$

Using equation (39), the estimated producer's accuracy of class j is computed by:

$$\widehat{PA}_j = \frac{\sum_{k=1}^{q} N_k^M \bar{y}_k}{\sum_{k=1}^{q} N_k^M \bar{x}_k} = \frac{N_1^M 0 + \cdots + N_{j-1}^M 0 + N_j^M \frac{n_{jj}}{n_{j+}} + N_{j+1}^M 0 + \cdots + N_q^M 0}{\sum_{k=1}^{q} N_k^M \frac{n_{kj}}{n_{k+}}} = \frac{N_j^M \frac{n_{jj}}{n_{j+}}}{\sum_{k=1}^{q} N_k^M \frac{n_{kj}}{n_{k+}}}$$

Note that $\sum_{k=1}^{q} N_k^M \frac{n_{kj}}{n_{k+}}$ represents the estimator for the real number of units in class j. Hence, the estimated producer's accuracy of class j is simplified to:

$$\widehat{PA}_j = \frac{N_j^M}{\widehat{N_j^R}} \frac{n_{jj}}{n_{j+}}$$

As Y is a dichotomous variable, equation (51), using equation (23), the sample variance of variable Y for the kth-stratum (i.e., kth-class) is given by:

$$s_{yk}^2 = \frac{n_{k+}}{n_{k+} - 1} \bar{y}_k (1 - \bar{y}_k) = \begin{cases} \frac{n_{k+}}{n_{k+} - 1} 0(1 - 0) & , if \text{ class } k \neq j \\ \frac{n_{j+}}{n_{j+} - 1} \frac{n_{jj}}{n_{j+}} \left(1 - \frac{n_{ii}}{n_{i+}}\right) & , if \text{ class } k = j \end{cases} = \begin{cases} 0 & , if \text{ class } k \neq j \\ \frac{n_{jj}}{n_{j+} - 1} \left(1 - \frac{n_{jj}}{n_{j+}}\right) & , if \text{ class } k = j \end{cases}$$

As X is a dichotomous variable, equation (52), using equation (24), the sample variance of variable X for the kth-stratum (i.e., kth-class) is given by:

$$s_{xk}^2 = \frac{n_{k+}}{n_{k+} - 1} \bar{x}_k (1 - \bar{x}_k) = \frac{n_{k+}}{n_{k+} - 1} \frac{n_{kj}}{n_{k+}} \left(1 - \frac{n_{kj}}{n_{k+}}\right) = \frac{n_{kj}}{n_{k+} - 1} \left(1 - \frac{n_{kj}}{n_{k+}}\right)$$

As Y and X are dichotomous variables, equations (50) and (51), respectively, using equation (25), the sample covariance between variables X and Y for the kth-stratum (i.e., kth-class) is given by:

$$
\begin{aligned}
s_{xyk} &= \frac{n_{k+}}{n_{k+}-1}\bar{y}_k(1-\bar{x}_k) \\
&= \begin{cases} \frac{n_{k+}}{n_{k+}-1}0\left(1-\frac{n_{kj}}{n_{k+}}\right) & ,if \text{ class } k \neq j \\ \frac{n_{j+}}{n_{j+}-1}\frac{n_{jj}}{n_{j+}}\left(1-\frac{n_{jj}}{n_{j+}}\right) & ,if \text{ class } k = j \end{cases} \\
&= \begin{cases} 0 & ,if \text{ class } k \neq j \\ \frac{n_{jj}}{n_{j+}-1}\left(1-\frac{n_{jj}}{n_{j+}}\right) & ,if \text{ class } k = j \end{cases} \\
&= s_{yk}^2
\end{aligned}
$$

Using equation (40), the estimated variance of the estimated producer's accuracy of class j is computed with:

$$
\begin{aligned}
\widehat{var}(\widehat{PA}_j) &= \frac{1}{\left(\sum_{k=1}^{q} N_k^M \bar{x}_k\right)^2}\sum_{k=1}^{q}\left(1-\frac{n_{k+}}{N_k^M}\right)\frac{(N_k^M)^2}{n_{k+}}\left[s_{yk}^2+\widehat{PA}_j^{\,2}s_{xk}^2-2\widehat{PA}_j s_{xyk}\right] \\
&= \frac{1}{\left(\sum_{k=1}^{q} N_k^M \frac{n_{kj}}{n_{k+}}\right)^2}\left[\widehat{PA}_j^{\,2}\sum_{k=1}^{q}\left(1-\frac{n_{k+}}{N_k^M}\right)\frac{(N_k^M)^2}{n_{k+}}s_{xk}^2+\sum_{k=1}^{q}\left(1-\frac{n_{k+}}{N_k^M}\right)\frac{(N_k^M)^2}{n_{k+}}\left(s_{yk}^2-2\widehat{PA}_j s_{xyk}\right)\right] \\
&= \frac{1}{\left(\sum_{k=1}^{q} N_k^M \frac{n_{kj}}{n_{k+}}\right)^2}\left[\widehat{PA}_j^{\,2}\sum_{k=1}^{q}\left(1-\frac{n_{k+}}{N_k^M}\right)\frac{(N_k^M)^2}{n_{k+}}s_{xk}^2+\sum_{k=1}^{q}\left(1-\frac{n_{k+}}{N_k^M}\right)\frac{(N_k^M)^2}{n_{k+}}\left(s_{yk}^2-2\widehat{PA}_j s_{yk}^2\right)\right] \\
&= \frac{1}{\left(\sum_{k=1}^{q} N_k^M \frac{n_{kj}}{n_{k+}}\right)^2}\left[\widehat{PA}_j^{\,2}\sum_{k=1}^{q}\left(1-\frac{n_{k+}}{N_k^M}\right)\frac{(N_k^M)^2}{n_{k+}}s_{xk}^2+\left(1-2\widehat{PA}_j\right)\sum_{k=1}^{q}\left(1-\frac{n_{k+}}{N_k^M}\right)\frac{(N_k^M)^2}{n_{k+}}s_{yk}^2\right] \\
&= \frac{1}{\left(\sum_{k=1}^{q} N_k^M \frac{n_{kj}}{n_{k+}}\right)^2}\left[\begin{matrix}\widehat{PA}_j^{\,2}\sum_{k=1}^{q}\left(1-\frac{n_{k+}}{N_k^M}\right)\frac{(N_k^M)^2}{n_{k+}}\frac{n_{kj}}{n_{k+}-1}\left(1-\frac{n_{kj}}{n_{k+}}\right)+ \\ +\left(1-2\widehat{PA}_j\right)\left(1-\frac{n_{j+}}{N_j^M}\right)\frac{\left(N_j^M\right)^2}{n_{j+}}\frac{n_{jj}}{n_{j+}-1}\left(1-\frac{n_{jj}}{n_{j+}}\right)\end{matrix}\right]
\end{aligned}
$$

Note that:

- $\sum_{k=1}^{q} N_k^M \frac{n_{kj}}{n_{k+}}$ is the estimated real number of units in class j, $\widehat{N_j^R}$ – equation (43);
- $\sum_{k=1}^{q}\left(1-\frac{n_{k+}}{N_k^M}\right)\frac{\left(N_k^M\right)^2}{n_{k+}}\frac{n_{kj}}{n_{k+}-1}\left(1-\frac{n_{kj}}{n_{k+}}\right)$ is the estimated variance of the estimated real number of units in class j, $\widehat{var}\left(\widehat{N_j^R}\right)$ – equation (44); and
- $\left(1-\frac{n_{j+}}{N_j^M}\right)\frac{1}{n_{j+}}\frac{n_{jj}}{n_{j+}-1}\left(1-\frac{n_{jj}}{n_{j+}}\right)$ is the estimated variance of the estimated user's accuracy of class j, $\widehat{var}\left(\widehat{UA}_i\right)$ – equation (49).

Therefore, the estimated variance of the estimated producer's accuracy of class j can be rewritten as:

$$
\widehat{var}(\widehat{PA}_j) = \frac{\widehat{PA}_j^{\,2}\widehat{var}(\widehat{N}_j^R)+(1-2\widehat{PA}_j)(N_j^M)^2\widehat{var}(\widehat{UA}_j)}{(\widehat{N}_j^R)^2}
$$

When the sample size per stratum is proportional to the stratum's area, i.e., the class are, we have that:

$$
\frac{N}{n} = \frac{N_K^M}{n_{k+}}
$$

and by equation (45), the estimated producer's accuracy of class j is computed by:

$$\widehat{PA}_j = \frac{N_j^M}{\widehat{N}_j^R}\frac{n_{jj}}{n_{j+}}$$
$$= \frac{N}{N\frac{n_{+j}}{n}}\frac{n_{jj}}{n}$$
$$= \frac{n_{jj}}{n_{+j}}$$

and the estimated variance of the estimated producer's accuracy of class j is computed by:

$$\widehat{var}(\widehat{PA}_j) = \frac{\widehat{PA}_j^{\,2}\,\widehat{var}\left(\widehat{N}_j^R\right) + \left(1 - 2\widehat{PA}_j\right)\left(N_j^M\right)^2 \widehat{var}\left(\widehat{UA}_j\right)}{\left(\widehat{N}_j^R\right)^2}$$
$$= \frac{\widehat{PA}_j^{\,2}\,\widehat{var}\left(\widehat{N}_j^R\right)}{\left(\widehat{N}_j^R\right)^2} + \frac{\left(1 - 2\widehat{PA}_j\right)\left(\frac{N}{n}n_{j+}\right)^2 \widehat{var}\left(\widehat{UA}_j\right)}{\left(N\frac{n_{+j}}{n}\right)^2}$$
$$= \frac{\widehat{PA}_j^{\,2}\,\widehat{var}\left(\widehat{N}_j^R\right)}{\left(\widehat{N}_j^R\right)^2} + \left(1 - 2\widehat{PA}_j\right)\left(\frac{n_{j+}}{n_{+j}}\right)^2 \widehat{var}\left(\widehat{UA}_j\right)$$

3.4 Estimation of the Overall Accuracy

The overall accuracy of a map is the proportion of units correctly classified, meaning the units that belong to the same class both on the map and in the reference. This metric and can be computed using equation (53):

$$OA = \frac{\sum_{u \in U} y_u}{N} \tag{53}$$

where N is the total number of units on the map and Y is a dichotomous variable that takes the value 1 when a unit belongs to the same class both in the reference and on the map, and 0 otherwise. That is:

$$Y(u) = \begin{cases} 1 & , if \text{ class } i \text{ is assigned to u in the reference } and\ the\ map \\ 0 & , otherwise \end{cases} \tag{54}$$

To estimate this quantity a mean estimator can be used, such as (13) or (36), respectively, when simple random samples or stratified random samples are considered

Note that, given the above definition of *Y*, equation (54), the sum of variable *Y* for all spatial units *u* represents the number of units that belong to class j both on the map and in the reference.

3.4.1 Simple random sampling

Given the considered definition of Y, equation (54), and using equation (1), the sample mean of the variable Y can be computed with:

$$\bar{y} = \frac{\sum_{l=1}^{n} y_l}{n}$$
$$= \frac{\sum_{k=1}^{q} n_{kk}}{n}$$

According to equation (13), the estimated overall accuracy is computed by:

$$\widehat{OA} = \bar{y}$$
$$= \frac{\sum_{k=1}^{q} n_{kk}}{n}$$

As Y is a dichotomous variable, equation (54), using equation (6), the sample variance of the variable Y is given by:

$$s_y^2 = \frac{n}{n-1}\bar{y}(1-\bar{y})$$
$$= \frac{n}{n-1}\frac{\sum_{k=1}^{q} n_{kk}}{n}\left(1 - \frac{\sum_{k=1}^{q} n_{kk}}{n}\right)$$
$$= \frac{\sum_{k=1}^{q} n_{kk}}{n-1}\left(1 - \frac{\sum_{k=1}^{q} n_{kk}}{n}\right)$$

The estimated variance of the estimated overall accuracy can be computed using equation (14). Therefore:

$$\widehat{var}\left(\widehat{OA}\right) = \left(1 - \frac{n}{N}\right)\frac{1}{n}s_y^2$$
$$= \left(1 - \frac{n}{N}\right)\frac{1}{n}\frac{\sum_{k=1}^{q} n_{kk}}{n-1}\left(1 - \frac{\sum_{k=1}^{q} n_{kk}}{n}\right)$$

When $n \ll N$, i.e., the population size is much larger than the sample size, $\left(1 - \frac{n}{N}\right) \approx 1$ and:

$$\widehat{var}\left(\widehat{OA}\right) = \frac{1}{n}\frac{\sum_{k=1}^{q} n_{kk}}{n-1}\left(1 - \frac{\sum_{k=1}^{q} n_{kk}}{n}\right)$$

3.4.2 Stratified random sampling based on map classes

Considering the definition of Y, equation (54), and using equation (18), the sample mean of variable Y for the kth-stratum (i.e., kth-class) is given by:

$$\bar{y}_k = \frac{\sum_{l=1}^{n_{k+}} y_{lk}}{n_{k+}}$$
$$= \frac{n_{kk}}{n_{k+}}$$

To estimate the overall accuracy equation (36) can be used, where the per strata mean estimator is computed with equation (33). Therefore and:

$$\widehat{OA} = \sum_{k=1}^{q} \frac{N_k^M}{N}\bar{y}_k$$
$$= \sum_{k=1}^{q} \frac{N_k^M}{N}\frac{n_{kk}}{n_{k+}}$$

As Y is a dichotomous variable, equation (54), using equation (23), the sample variance of the variable Y of kth-stratum (i.e., kth-class) is given by:

$$s_{yk}^2 = \frac{n_{k+}}{n_{k+}-1}\bar{y}_k(1-\bar{y}_k)$$
$$= \frac{n_{k+}}{n_{k+}-1}\frac{n_{kk}}{n_{k+}}\left(1 - \frac{n_{kk}}{n_{k+}}\right)$$
$$= \frac{n_{kk}}{n_{k+}-1}\left(1 - \frac{n_{kk}}{n_{k+}}\right)$$

Note that $\frac{n_{kk}}{n_{k+}}$ represents the estimator for the user's accuracy of class k. Hence, the estimated overall accuracy can be rewritten to:

$$\widehat{OA} = \frac{1}{N}\sum_{k=1}^{q} N_k^M \widehat{UA}_k$$

Using equations (34) and (37), the estimated variance of the estimated overall accuracy is given by:

$$\begin{aligned}\widehat{var}\big(\widehat{OA}\big) &= \sum_{k=1}^{q}\left(1-\frac{n_{k+}}{N_k^M}\right)\left(\frac{N_k^M}{N}\right)^2 \frac{1}{n_{k+}} s_{yk}^2 \\ &= \frac{1}{N^2}\sum_{k=1}^{q}\left(1-\frac{n_{k+}}{N_k^M}\right)(N_k^M)^2 \frac{1}{n_{k+}}\frac{n_{kk}}{n_{k+}-1}\left(1-\frac{n_{kk}}{n_{k+}}\right)\end{aligned}$$

Note that, by equation (49), $\left(1-\frac{n_{k+}}{N_k^M}\right)\frac{1}{n_{k+}}\frac{n_{kk}}{n_{k+}-1}\left(1-\frac{n_{kk}}{n_{k+}}\right)$ is the estimated variance of user's accuracy of class k, $\widehat{var}\big(\widehat{UA}_k\big)$. Hence, the estimated variance of the estimated overall accuracy can be rewritten as:

$$\widehat{var}\big(\widehat{OA}\big) = \frac{1}{N^2}\sum_{k=1}^{h}\left(N_k^M\right)^2 \widehat{var}\big(\widehat{UA}_k\big)$$

When the sample size is proportional to the strata area, i.e., the classes' area, we have:

$$\frac{N}{n} = \frac{N_k^M}{n_{k+}}$$

In this case the estimated overall accuracy is computed with:

$$\begin{aligned}\widehat{OA} &= \sum_{k=1}^{q}\frac{N_k^M}{N}\frac{n_{kk}}{n_{k+}} \\ &= \sum_{k=1}^{q}\frac{N}{N}\frac{n_{kk}}{n} \\ &= \frac{1}{n}\sum_{k=1}^{q} n_{kk}\end{aligned}$$

and the estimated variance of the estimated overall accuracy is given by:

$$\begin{aligned}\widehat{var}\big(\widehat{OA}\big) &= \sum_{k=1}^{q}\left(\frac{N_k^M}{N}\right)^2 \widehat{var}(\widehat{UA}_k) \\ &= \sum_{k=1}^{q}\left(\frac{n_{k+}}{n}\right)^2 \widehat{var}(\widehat{UA}_k) \\ &= \frac{1}{n^2}\sum_{k=1}^{q}(n_{k+})^2\, \widehat{var}(\widehat{UA}_k)\end{aligned}$$

4 Final Remarks

With technological advancements and the increasing availability of Earth observation data, thematic maps, such as LULC maps, are widely used for decision making in a broad scope of applications. However, the reliability of decisions made based on such maps is very much dependent of the map's quality. Therefore, accurate accuracy assessment in increasingly important.

The technical report demonstrates how the formulas that are frequently used to estimate the classes true area and the maps thematic accuracy metrics are derived from statistics metrics, and what are the considered assumptions that generate such formulas.

The methods discussed in this work can also be applied to other fields of study, with the necessary adaptation of terminology from remote sensing to the relevant field. For instance, in machine learning, producer's accuracy corresponds to sensitivity, and user's accuracy corresponds to positive predictive value. Similarly, the metrics used in remote sensing can be estimated in analogous ways across various disciplines. This approach is useful wherever there is a need to assess thematic accuracy, enabling a seamless transfer of accuracy assessment methods between fields.

Acknowledgements

This work was supported by a doctoral grant from FCT – Fundação para a Ciência e a Tecnologia, I.P. (Grant No. 2022.09754.BD). It was also partially financed through national funds by FCT, within the framework of the Project UIDB/00326/2025 and UIDP/00326/2025. In addition, this work is funded by FCT/MEC through national funds and co-funded by the FEDER-PT2020 partnership agreement under the project UIDB/00308/2020.

References

- Cochran, W.G., 1977. Sampling Techniques, Wiley Series in Probability and Statistics. Wiley.

- Fonte, C.C., See, L., Laso-Bayas, J.C., Lesiv, M., Fritz, S., 2020. Assessing the accuracy of land use land cover (lulc) maps using class proportions in the reference data, in: ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences. Presented at the XXIV ISPRS Congress, Commission III (Volume V-3-2020) - 2020 edition, Copernicus GmbH, pp. 669–674. https://doi.org/10.5194/isprs-annals-V-3-2020-669-2020

- Olofsson, P., Foody, G.M., Herold, M., Stehman, S.V., Woodcock, C.E., Wulder, M.A., 2014. Good practices for estimating area and assessing accuracy of land change. Remote Sensing of Environment 148, 42–57. https://doi.org/10.1016/j.rse.2014.02.015

- Stehman, S.V., Foody, G.M., 2019. Key issues in rigorous accuracy assessment of land cover products. Remote Sensing of Environment 231, 111199. https://doi.org/10.1016/j.rse.2019.05.018

- Thompson, S.K., 2012. Sampling, 3rd ed. ed, Wiley series in probability and statistics. Wiley, Hoboken, N.J.

Appendix A

Table 3. Estimators and their estimated variances for thematic accuracy metrics (rows: true number of units in a class, user's accuracy, producer's accuracy, and overall accuracy) under different sampling approaches (columns: simple random sampling; stratified random sampling based on map classes; and stratified random sampling based on map classes with proportional allocation according to class size).

	Simple random sampling	**Stratified random sampling based on map classes**	**Stratified random sampling based on map classes, with the sample proportionally allocated according to the size of each class**
True Number of Units in a Class	$\widehat{N}_j^R = N\frac{n_{+j}}{n}$	$\widehat{N}_j^R = \sum_{k=1}^{q} N_k^M \frac{n_{kj}}{n_{k+}}$	$\widehat{N}_j^R = N\frac{n_{+j}}{n}$
	$\widehat{var}(\widehat{N}_j^R) = \frac{\widehat{N}_j^R(N - \widehat{N}_j^R)}{n-1}$	$\widehat{var}(\widehat{N}_j^R) = \sum_{k=1}^{q} \frac{(N_k^M)^2}{n_{k+}-1}\frac{n_{kj}}{n_{k+}}\left(1 - \frac{n_{kj}}{n_{k+}}\right)$	$\widehat{var}(\widehat{N}_j^R) = \left(\frac{N}{n}\right)^2 \sum_{k=1}^{q} \frac{n_{k+}}{n_{k+}-1} n_{kj}\left(1 - \frac{n_{kj}}{n_{k+}}\right)$
User's Accuracy of a Class	$\widehat{UA}_i = \frac{n_{ii}}{n_{i+}}$	$\widehat{UA}_i = \frac{n_{ii}}{n_{i+}}$	$\widehat{UA}_i = \frac{n_{ii}}{n_{i+}}$
	$\widehat{var}(\widehat{UA}_i) = \frac{\widehat{UA}_i(1-\widehat{UA}_i)}{n_{i+}}$	$\widehat{var}(\widehat{UA}_i) = \frac{\widehat{UA}_i(1-\widehat{UA}_i)}{n_{i+}-1}$	$\widehat{var}(\widehat{UA}_i) = \frac{\widehat{UA}_i(1-\widehat{UA}_i)}{n_{i+}-1}$
Producer's Accuracy of a Class	$\widehat{PA}_j = \frac{n_{jj}}{n_{+j}}$	$\widehat{PA}_j = \frac{N_j^M}{\widehat{N}_j^R}\frac{n_{jj}}{n_{j+}}$	$\widehat{PA}_j = \frac{n_{jj}}{n_{+j}}$
	$\widehat{var}(\widehat{PA}_j) = \frac{\widehat{PA}_j(1-\widehat{PA}_j)}{n_{+j}}$	$\widehat{var}(\widehat{PA}_j) = \frac{\widehat{PA}_j^{\,2}\,\widehat{var}(\widehat{N}_j^R) + (1-2\widehat{PA}_j)(N_j^M)^2\,\widehat{var}(\widehat{UA}_j)}{(\widehat{N}_j^R)^2}$	$\widehat{var}(\widehat{PA}_j) = \frac{\widehat{PA}_j^{\,2}\,\widehat{var}(\widehat{N}_j^R)}{(\widehat{N}_j^R)^2} + (1-2\widehat{PA}_j)\left(\frac{n_{j+}}{n_{+j}}\right)^2 \widehat{var}(\widehat{UA}_j)$
Overall Accuracy	$\widehat{OA} = \frac{\sum_{k=1}^{q} n_{kk}}{n}$	$\widehat{OA} = \frac{1}{N}\sum_{k=1}^{q} N_k^M \widehat{UA}_k$	$\widehat{OA}\& = \frac{\sum_{k=1}^{q} n_{kk}}{n}$
	$\widehat{var}(\widehat{OA}) = \frac{\widehat{OA}}{n-1}(1-\widehat{OA})$	$\widehat{var}(\widehat{OA}) = \frac{1}{N^2}\sum_{k=1}^{q}(N_k^M)^2\,\widehat{var}(\widehat{UA}_k)$	$\widehat{var}(\widehat{OA}) = \frac{1}{n^2}\sum_{k=1}^{q}(n_{k+})^2\,\widehat{var}(\widehat{UA}_k)$