applied sciences

MDPI

Article

Thematic Accuracy Quality Control by Means of a Set of Multinomials

Francisco J. Ariza-López [1], José Rodríguez-Avi [2], María V. Alba-Fernández [2] and José L. García-Balboa [1,*]

[1] Departamento de Ingeniería Cartográfica, Geodésica y Fotogrametría, Universidad de Jaén, 23071 Jaén, Spain; fjariza@ujaen.es

[2] Departamento de Estadística e Investigación Operativa, Universidad de Jaén, 23071 Jaén, Spain; jravi@ujaen.es (J.R.-A.); mvalba@ujaen.es (M.V.A.-F.)

* Correspondence: jlbalboa@ujaen.es; Tel.: +34-953-212844

Received: 3 September 2019; Accepted: 9 October 2019; Published: 11 October 2019

Abstract: The error matrix has been adopted as both the "de facto" and the "de jure" standard way to report on the thematic accuracy assessment of any remotely sensed data product. This perspective assumes that the error matrix can be considered as a set of values following a unique multinomial distribution. However, the assumption of the underlying statistical model falls down when true reference data are available for quality control. To overcome this problem, a new method for thematic accuracy quality control is proposed, which uses a multinomial approach for each category and is called QCCS (quality control column set). The main advantage is that it allows us to state a set of quality specifications for each class and to test if they are fulfilled. These requirements can be related to the percentage of correctness in the classification for a particular class but also to the percentage of possible misclassifications or confusions between classes. In order to test whether such specifications are achieved or not, an exact multinomial test is proposed for each category. Furthermore, if a global hypothesis test is desired, the Bonferroni correction is proposed. All these new approaches allow a more flexible way of understanding and testing thematic accuracy quality control compared with the classical methods based on the confusion matrix. For a better understanding, a practical example of an application is included for classification with four categories.

Keywords: thematic accuracy; quality control column set; multinomial test

1. Introduction

Spatial data products (e.g., cadaster, road networks, digital elevation models, land cover, fire and drought incidence maps, etc.) are of great importance for environment modeling, decision making, climate change assessment, and so on. They, therefore, have great economic aggregated value [1]. However, spatial data are not exempt from errors, so spatial data sets represent a significant source of error in any analysis that uses them as input [2]. For this reason, spatial data quality has been a major research concern in previous decades [3]. Spatial data quality has several perspectives (e.g., accessibility, usability, integrity, interoperability, etc.), and several quality elements (completeness, logical consistency, thematic accuracy, positional accuracy, temporal accuracy) have been defined by the international standard ISO 19113 [4]. Accuracy comprises trueness and precision, as stated by ISO 5725-1 [5]. Trueness is the absence of bias, and precision is a measure of dispersion. In thematic mapping, dealing with categories, the term accuracy means trueness, that is, the degree of "correctness" of a map classification. A classification may be considered accurate if it provides an unbiased representation of the reality (agrees with reality) or conforms to the "truth". Thematic accuracy is defined by ISO 19157 [6] as the accuracy of quantitative attributes and the correctness of non-quantitative attributes

and of the classifications of features and their relationships. Classification correctness is defined by the same standard as the comparison of the classes assigned to features or their attributes to a universe of discourse (e.g., ground truth or reference data) [6]. Classification correctness is the main concern in any remote-sensed-derived product (e.g., land cover, fire and drought incidence maps, etc.) and, in general, for any kind of spatial data (e.g., vector data, such as cadastral parcels, road networks, topographic databases, etc.). The evaluation of thematic accuracy is a subject of interest, proof of this being the existence of numerous references, for instance [7] for ecosystem maps, [8] for land change maps, [9] for vegetation inventory, and [10] for land cover.

The main components of a thematic accuracy assessment are [11] (i) the sampling design used to select the reference sample; (ii) the response design used to obtain the reference land-cover classification for each sampling unit; and (iii) the estimation and analysis procedures. This method has been widely adopted (e.g., [12,13]), and also extended; for example, [8] proposed a method with eight steps as a good practice for assessing the accuracy of land change, in which some steps are coincident with those used in [11]. However, for a proper classification correctness assessment, a classification scheme is also needed. A classification scheme has two critical components [10]: (i) a set of labels and (ii) a set of rules for assigning labels so that a unique assignation of classes is possible. This is because classes (labels) are mutually exclusive (there are no overlaps) and totally exhaustive with respect to the thematic universe of discourse (there are no omissions or commissions of classes).

From our point of view, the two previous aspects must be considered from a more general perspective of the production processes of spatial data. From this perspective, the first thing to consider is a specification of the product (e.g., in the sense of ISO 19131, [14]). This specification should contain the classification scheme but also a specification of the level of quality required for each category (e.g., at least 90% of the classification correctness for category A) and the grade of confusion allowed between categories (e.g., at most, 5% confusion between categories A and B). These quality grades must be in accordance with the processes' voice (capacity to give some quality grade) and the user's voice (quality needs for a specific use case). Quality is expressed by means of measures or indices, and those must be well-defined and defined prior to data production. For instance, ISO 19157 [6] offers a list of twelve components (identifier, name, definition, etc.) for defining a standardized measure but also a complete set of standardized measures (see Annex D of ISO 19157), where some of them can be used for thematic accuracy assessment (e.g., measures from #60 to #64 for classification correctness). In addition, a data product specification must also include the quality evaluation method [15] in order to assure standardization of the assessment procedures. Here, the three main components for a thematic accuracy assessment stated by [11] can be considered. Finally, a standardized meta-quality assessment is desirable in order to have actual confidence about the quality of the quality assessment [15]. The international standard ISO 19157 promotes this new quality element for spatial data products. This idea was also presented by [11] when they pointed out the desirability of using a quality control measure to evaluate the accuracy of the reference land-cover classifications used for quality control.

In [16], four major historical stages in thematic accuracy assessment are identified: (i) the "it looks good" age, (ii) the "non-site-specific assessment" age, (iii) the "site-specific assessment" age, and (iv) the "error matrix" age. The confusion matrix is currently at the core of the accuracy assessment literature [17] and, as stated by [18], the error matrix has been adopted as both the "de facto" and the "de jure" standard—the way to report on the thematic accuracy of any remotely-sensed-data product (e.g., image-derived data). Of course, the same tool can be used for any kind of data that directly originates in vector form.

A confusion matrix and the indices derived from it are statistical tools used for the analysis of paired observations. When the objective is to compare two classified data products (by different processes, different operators, different times, or something similar), the observed frequencies in a confusion matrix are assumed to be modeled by a multinomial distribution (forming a vector after ordering by columns, for instance). The indexes derived, like overall accuracy, kappa, producer's and

user's accuracies, and so on, are based on this assumption (multinomial distribution), and they make sense due to the complete randomness of the elements inside the confusion matrix. However, when true reference data are available, the inherent randomness falls down, therefore also the assumption of the underlying statistical model. Suppose the reference data are located by column. If the reference data are considered as the truth, the total number of elements we know that belong to a particular category can be correctly classified or confused with other categories, but they will always be located in the same column but never in other different columns (category). This fact implies that the inherent randomness of the multinomial is not possible now. However, we can deal with the available classification by considering a multinomial distribution for each category (column) instead of the initial multinomial distribution which involved all the elements in the matrix. For this reason, we call this approach the quality control column set (QCCS). Therefore, the goal of this study is to develop the statistical basis of this new approach and to give an example of its application. The statistical foundation rests both on a multinomial approach to each column of the confusion matrix and on an exact statistical test. The paper is organized as follows: Section 2 defines the new approach and compares it with the standard use of confusion matrices; Section 3 is devoted to presenting the proposed method by using some hypothetical examples; in Section 4, an actual example is analyzed and discussed. Finally, in Section 5, some conclusions are stated.

2. Quality Control Column Set

In this section, we first present an approximation to what a confusion matrix is and, subsequently, we present the concept of QCCS and the difference between the two approaches. The aspect that differentiates both approaches—the quality of the reference data—is highlighted.

A confusion matrix, or error matrix, is a contingency table, which is a statistical tool for the analysis of paired observations. The confusion matrix is proposed and defined as a standard quality measure for spatial data (measure #62) by ISO 19157 [6] in Annex D with the name "misclassification matrix" or "error matrix". For a given geographical space, the content of a confusion matrix is a set of values accounting for the degree of similarity between paired observations of k classes in a controlled data set (CDS), and the same k classes of a reference data set (RDS):

$$\text{CM(i,j)} = [\#\text{items of class (j) of the RDS classified as class (i) in the CDS}]. \quad (1)$$

As indicated by the study presented in [19], the most frequent number of classes k is between 3 and 7, although there are matrices that reach up to 40 categories. Usually, the RDS and CDS are located by columns and by rows, respectively. So, it is a $k \times k$ squared matrix. The diagonal elements of a confusion matrix contain the number of correctly classified items in each class or category, and the off-diagonal elements contain the number of confusions. So, a confusion matrix is a type of similarity assessment mechanism used for thematic accuracy assessments.

A confusion matrix is not free of errors ([17,20]), and for this reason, a quality assurance of intervening processes is needed, e.g., the proposal of [11] can be considered in this way (in order to apply a statistically rigorous accuracy assessment). As pointed out by [21], obtaining a reliable confusion matrix is a weak link in the accuracy assessment chain. Here, a key element is the RDS, denoted sometimes as the "ground truth", which can be totally inappropriate and, in some cases, very misleading [10] and should be avoided ([22,23]). As pointed out by several studies ([24–27]), the RDS often contains errors and sometimes possibly more errors than the CDS. Here, the major problem comes from the fact that classifications are often based on highly subjective interpretations [28]. The problem with a lack of quality in the reference data is still current [29], and the thematic quality of products derived from remote sensing still presents problems [30]. We understand that this situation is due to the fact that, in most cases, the RDS is simply another set of data (just another classification) and not a true reference (error-free or of better quality).

The above-mentioned situation does not occur in the quality assessment of other components of spatial data quality; in this way, as indicated by [31], compared to positional accuracy, there is a clear lack of standardization. For example, in the case of positional accuracy, the ASPRS standard [32] establishes the following requirement: "The independent source of higher accuracy for checkpoints shall be at least three times more accurate than the required accuracy of the geospatial data set being tested". This situation is directly achievable when working with topographic and geodetic instruments, but it is not directly attainable when working with thematic categories because of the high subjectivity of interpretations ([28,33]). However, we believe that this situation should guide all processes for determining the RDS of an assessment of thematic accuracy. In this sense, as an example of good practices in accuracy assessment, [8] the higher quality of the sample is included as a fundamental premise.

RDSs are usually derived from ground observations or from image interpretations. Fieldwork derived from ground observation is the most accurate possible source for thematic accuracy assessment, but it may be cost-prohibitive for many projects. For this reason, common image interpretations are preferred. In both cases, the subjectivity of interpretations is present, and operators' errors are expected, as was recently analyzed by [33] for the case of visual interpreters. Therefore, in order to actually achieve greater accuracy for the RDS, some quality assurance actions need to be deployed in order to reduce the subjectivity of the interpretations. For instance, [34] proposed (i) using a group of selected operators, (ii) designing a specific training procedure for the group of operators in each specific quality control (use case), (iii) calibrating the work of the group of operators in a controlled area, (iv) supplying the group with good written documentation of the product specifications and the quality control process, (v) helping the group with good service support during the quality-control work and socializing the problems and the solutions and, finally, (vi) proceeding to classification based on a multiple assignation process produced by the operators of the group, achieving agreements where needed. The same idea of obtaining more cohesive groups of experts for validation is present in [35], where it is indicated that you have to get a group with common criteria, and, to this end, a five-day workshop should be held. In relation to the achievement of agreement, [36] proposes that validation sampling units be reviewed by nine experts and that the adoption of a label requires a consensus of at least 6 out of 9 among these experts. All of these actions are quality assurance actions and must be deployed, paying special attention to improving trueness (reducing systematic differences between operators and reality), precision (increasing agreement between operators in each case), and uniformity (increasing the stability of operators' classifications under different scenarios).

The accuracy of the RDS is of great relevance when considering the statistical tools of thematic quality control. If the RDS does not have the quality to be a reference, the confusion matrix can be understood as a complete multinomial, or an almost complete multinomial if structural zeros are present. From this perspective, the analyses based on the confusion matrix are correct (e.g., overall accuracy, kappa, users' and producers' accuracies, and so on). Thus, if the RDS does have the quality to be a reference, it is not correct to work with the complete confusion matrix because the inherent randomness in the matrix falls down. However, we can manage the data under a new approach that consists of separating the matrix into columns (one for each category) and redefining a multinomial distribution for each category (column). This new idea helps us to clearly differentiate this situation from the previous one based on using the confusion matrix as a complete multinomial. Of course, the thematic quality control data may continue to be displayed as a table (matrix), but the analysis should be carried out by columns.

Within this new approach, we propose a method of category-wise control that allows the statement of our preferences of quality, category by category, but also the statement of misclassifications or confusion among classes. These preferences are expressed in terms of the minimum percentages required for well-classified items and the maximum percentage allowed for misclassifications among classes within each column.

In order to illustrate the application of the above with an example, Figure 1 shows a confusion matrix from [37] with the results from the accuracy assessment of the classification of a synthetic data set with four categories. Now, let us consider that the RDS used in this assessment does have the quality to be a reference. Therefore, the data from Figure 1 cannot be understood as a complete multinomial but, rather, as a set of four multinomials, one for each category (column). Figure 2 illustrates this fact with locks that symbolize that the marginals of the columns are fixed and therefore, the new structure "quality control column set" (QCCS) has to be considered instead of the classical method based on the confusion matrix.

		RDS			
		Wo	**G**	**N**	**Wa**
CDS	**Wo**	47	3	0	0
	G	4	40	6	0
	N	0	5	45	0
	Wa	0	0	2	48

Figure 1. Example of the confusion matrix [37]. Wo = Woodland, G = Grassland, N = Non-vegetated, Wa = Water.

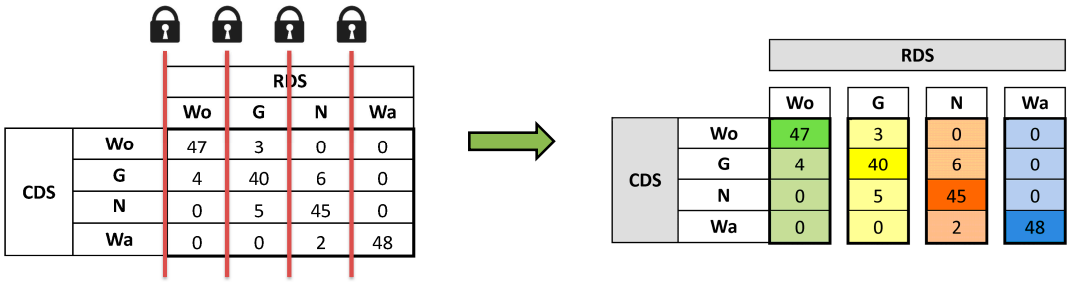

Figure 2. The new structure called the "quality control column set" (QCCS), applied to the data from Figure 1. The locks symbolize that the marginals of the columns are fixed. For clarity, each column is presented in a different color, highlighting the number of correctly classified items.

Following with the example, once the QCCS structure is considered, our proposal allows us to consider a set of quality specifications in the following manner: For each category, a classification level can be stated, but also a misclassification level with each other category (or groups of them). In Table 1, we summarize an example of quality specifications that are supposed for the synthetic data presented in [37]. We indicate, for each category, the minimum percentage required for well-classified items, but also the maximum percentage allowed in misclassifications. As can be seen, we assume high percentages of well-classified items (≥95%, ≥88%, ≥90%, and ≥99%), low percentages in some misclassifications (≤4%, ≤10%, and ≤8%) and some other almost non-existent misclassifications (≤1% and ≤2%). Several misclassification levels group two or more categories. This possibility of merging categories offers a more flexible quality control analysis. For example, in Table 1, it is indifferent to the expert that the water is misclassified with any of the remaining three classes, but in relation to the *Grassland* class, the expert is more worried about a misclassification with *Woodland* or *Water* (≤2%) than a misclassification with *Non-vegetated* (≤8%).

As will be seen in later sections, our proposal of an exact multinomial test requires an order in the probabilities. Therefore, the set of quality specifications of Table 1 is not complete until we state, for each category, which order is assumed for them. For our example, we assume the order in which specifications are presented in Table 1. That is to say, for the category *Woodland*, the specification SPWo#1 is the most important one to be fulfilled, the second is SpWo#2, and finally, the third is SpWo#3, but any other order could have been considered.

Table 1. Example of a set of specifications [1]: The quality levels required for each category and the percentage of misclassifications allowed between classes within each category.

Category	Specification ID	Description
Woodland	SpWo#1	95% of the minimum percentage required in well-classified items (≥95%)
	SpWo#2	4% of the maximum percentage allowed in misclassifications with *Grassland* (≤4%)
	SpWo#3	1% of the maximum percentage allowed in misclassifications with both *Non-vegetated* and *Water* (≤1%)
Grassland	SpG#1	88% of the minimum percentage required in well-classified items (≥88%)
	SpG#2	10% of the maximum percentage allowed in misclassifications with *Non-vegetated* (≤10%)
	SpG#3	2% of the maximum percentage allowed in misclassifications with both *Woodland* and *Water* (≤2%)
Non-vegetated	SpN#1	90% of the minimum percentage required in well-classified items (≥90%)
	SpN#2	8% of the maximum percentage allowed in misclassifications with *Grassland* (≤8%)
	SpN#3	2% of the maximum percentage allowed in misclassifications with both *Woodland* and *Water* (≤2%)
Water	SpWa#1	99% of the minimum percentage required in well-classified items (≥99%)
	SpWa#2	1% of the maximum percentage allowed in misclassifications with the other categories (≤1%)

[1] These specifications are only examples.

3. QCCS Statistical Basis

The core of this section is twofold: (i) we introduce the QCCS structure starting from the confusion matrix and point out the implications of RDS as a reference for dealing with the data from a statistical point of view, and (ii) we state the quality levels required for the classification and determine when we will decide whether such quality levels are achieved or not by means of an exact multinomial test.

3.1. From the Confusion Matrix to the Set of Multinomial Distributions

Let $\Gamma_1, \ldots, \Gamma_k$ be the k true categories of the RDS and $G_1, \ldots, G_k$ be the categories of the CDS, assigned by the classification process. Adscription of an item into the cell (i, j) implies that it belongs to category j in the RDS, and, by the classification procedure, it is classified as corresponding to category i in the CDS. So, the elements inside the cells (i, j), $i = j$ are correctly classified, whereas an assignation into the cell (i, j), $i \neq j$ implies an error or misclassification. Finally, n_{ij} indicates the number of items assigned to the cell (i, j), and n_{+j} is the sum of items that belong to the true category Γ_j, $j = 1, \ldots, k$. Table 2 displays the disposition of classification results corresponding to a $k \times k$ cross-classification.

Table 2. Notation of a $k \times k$ cross-classification. CDS: controlled data set; RDS: reference data set.

		RDS			
		$\boldsymbol{\Gamma_1}$	$\boldsymbol{\Gamma_2}$	$\ldots$	$\boldsymbol{\Gamma_k}$
CDS	$\boldsymbol{G_1}$	n_{11}	n_{12}	$\ldots$	n_{1k}
	$\boldsymbol{G_2}$	n_{21}	n_{22}	$\ldots$	n_{2k}
	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
	$\boldsymbol{G_k}$	n_{k1}	n_{k2}	$\ldots$	n_{kk}
Total		n_{+1}	n_{+2}		n_{+k}

If we consider the independence and randomness in the sampling procedure, the first approach is to see the whole matrix as one realization of a multinomial distribution. Recall that the multinomial distribution arises from an extension of the binomial distribution to situations where each trial has

more than two possible outcomes. In a classical confusion matrix, the vector of possible success is one obtained from the cross-classification derived from the classification procedure. To be rigorous, if we realize n independent trials where we carry out the cross-classification of an item in one of $c = k \times k$ cells, we define X_i, $i = 1, \ldots, c$ as the number of times that the outcome associated with cell i occurs after reordering the matrix by columns to construct a vector. So, the vector $X = (X_1, X_2, \ldots, X_c)$ is said to have a multinomial distribution with parameter n, the number of trials, and the probability vector $\pi = (\pi_1, \ldots, \pi_c)$, with each $\pi_i \geq 0$ corresponding to the occurrence probability of the classification results in each cell, satisfying $\sum_{i=1}^{c} \pi_i = 1$, in short, $M(n, \pi)$. Hence, the mass probability function of this distribution $M(n, \pi)$ is given by

$$P[(X_1 = n_1, \ldots, X_c = n_c) = \frac{n!}{n_1! \ldots n_c!} \pi_1^{n_1} \ldots \pi_c^{n_c}, \quad (2)$$

where $\sum_{i=1}^{c} n_i = n$.

However, if the RDS has the quality to be a reference, not all possible values of this multinomial are feasible. Next, we point out two situations as illustrative examples:

1. For any category Γ_j (reference), there are n_{+j} items in this column, so it is impossible to find a value $n_{ij} > n_{+j}$ in the same column. This could be possible only if RDS is not considered as a reference, that is, without constraints in the randomness of the matrix. In addition, the n_{+j} elements can be correctly classified or confused with other categories, but they will always be located in the same column and never in other different columns (categories). This implies that the inherent randomness of the multinomial is not possible now.
2. According to Equation (2), a possible outcome of an $M(n, \pi)$ can concentrate all items in one cell, say $(n, 0, \ldots, 0)$, but if RDS is considered to be a reference, this outcome will never be observed, because the values n_{+j} are always present and, in consequence, all outcomes are partially grouped into subsets that sum n_{+j}.

Therefore, as stated in Section 2, when RDS is considered as a reference, the values n_{+j} are fixed, and by redefining the outcomes for this number of trials, we can obtain a new multinomial distribution, one for each category.

So, we take advantage of this fact and we propose to deal with the information contained in Figure 1 as the new structure called the "quality control column set" (QCCS) (Figure 2) in order to differentiate it from the classical confusion matrix concept. Now, each column in the QCCS structure can be modeled as a different multinomial distribution. Moreover, a column can be modeled as a binomial distribution if a single misclassification level is set for all other $(k - 1)$ categories in the specifications (e.g., specification SpWa#2 in Table 1).

The main advantage of this approach is that it allows us to assume specific quality requirements for each class analyzed and to test them. Now, we can propose a category-wise control method that allows us to establish levels of quality, category by category (column by column), related to the percentage of correctness in the classification for a particular class and to the percentage of possible misclassifications or confusions between classes. So, for each category and in a separate way, we can state a set of quality specifications, and in order to test whether such specifications are achieved or not, we propose an exact multinomial test.

3.2. Thematic Accuracy Quality Control

In order to fix the terms to be used, for a particular category Γ, we call the class that gives its name the "main class", and the "rest of classes" are the remaining initial classes in the confusion matrix or new ones after the merging of one or more initial classes.

We have to put the rest of the classes in an assumed order, which we can establish freely. As an example, we come back to Table 1. In the quality requirements for the *Grassland* category, the main class

is *Grassland* and, as stated at the end of Section 2, a confusion with *Non-vegetated* is considered more important than a confusion with both *Woodland* and *Water*. Therefore, the second class is *Non-vegetated*, and the third class is the new class *Woodland + Water*. So, we work with a multinomial distribution with three classes, $X = (X_1, X_2, X_3)$ where X_1 represents the number of items in the main class that are correctly classified, X_2 counts the number of confusions between the main class and the second class, and X_3 stands for the number of misclassifications between the main class and the third class.

According to the previous notation, quality control involves establishing the following quality conditions for each category:

- The minimum proportion of well-classified items in the main class;
- The maximum possible $k-1$ misclassification proportions between the main class and the rest of the classes.

At this point, we have to define when the classification results of a category agree with the specifications given previously. The criteria adopted are that the specifications are not fulfilled when:

i. The number of correct items in the main class is lower than those expected, or
ii. If the number of correct items in the main class is equal to those expected, the number of misclassifications with the second class is greater than the expected, or
iii. If the number of correct items and the number of misclassifications with the second class is equal to those expected, the number of misclassifications with the third class is greater than that expected, and so on.

As a consequence, thematic accuracy quality control implies the proposal of a test for testing the following null hypothesis:

$$\mathbb{H}_0\text{: The classification results agree with the set of specifications stated in the quality control for each category.} \tag{3}$$

Assuming the QCCS is an independent set of k multinomial distributions, we can propose 1, 2, or even k different null hypotheses (one for each category/column), and hence, we can perform up to k independent tests. In the case of desiring a global hypothesis test, the Bonferroni correction is used. Therefore, this approach allows a more flexible way of understanding and testing thematic accuracy quality control than the classical methods based on the confusion matrix.

3.3. The Exact Multinomial Test

From now on, in order to maintain the notation as simple as possible, we are going to test the null hypothesis of Equation (3) for a fixed category, say Γ, that will represent any of the set $\{\Gamma_1, \Gamma_2, \ldots, \Gamma_k\}$, and hence we will test whether the thematic quality specifications previously stated are achieved or not in the sense of the null hypothesis $\mathbb{H}_0$.

Thus, we consider that for a particular category Γ, we have m elements to classify (which corresponds to its marginal value in Table 2), and we define X_1 as the number of elements correctly classified in the main class Γ, and we define $X_i, i = 2, \ldots, q, q \leq k$ as the number of misclassifications between the main class and the rest of the classes in decreasing order, assumed in the misclassification levels. In this way, we obtain a multinomial vector $X = \left(X_1, X_2, \ldots, X_q\right)$ with the parameter m and the probability vector $p_0 = \left(p_{10}, p_{20}, \ldots, p_{q0}\right)$, $p_{i0} \geq 0$, $\sum_{i=1}^{q} p_{i0} = 1$, with $p_{10} \geq p_{20} \geq \ldots \geq p_{q0}$, with p_{i0} being the percentage stated in the limits expressed in a set of specifications (e.g., Table 1) under the null hypothesis.

To test Equation (3), an exact test is proposed. Remember that an exact test calculates the empirical probability of obtaining an outcome as different from the null hypothesis as to the outcome observed in the data. So, the p-value is computed by adding up the probabilities of feasible outcomes in $M(m, p_0)$ (say, $P[X = x]$) under the alternative hypothesis, starting from the observed classification results.

To decide when a classification result $X = \left(x_1, x_2, \ldots, x_q\right)$ is worse than $Y = \left(y_1, y_2, \ldots, y_q\right)$ according to the specifications (denoted by $X <_m Y$), we consider the following partial order relation between X and Y in the set $\left\{x_1, x_2, \ldots, x_q \in \{0, 1, \ldots, m\} : \sum_{i=1}^{q} x_i = m\right\}$, which is similar to the lexicographic order, in the following sense,

$$\begin{gathered} X <_m Y \text{ if } (x_1 < y_1) \vee \\ \{(x_1 = y_1) \wedge (x_2 > y_2)\} \vee \\ \{(x_1 = y_1) \wedge (x_2 = y_2) \wedge (x_3 > y_3)\} \ldots\ldots\ldots \\ \left\{(x_1 = y_1) \wedge \ldots \wedge \left(x_{q-1} = y_{q-1}\right)\right\} \wedge \left(x_q > y_q\right), \end{gathered} \tag{4}$$

where $\wedge$ and $\vee$ represent the basic operators of Boolean algebra "AND" and "OR", respectively.

$X = \left(x_1, x_2, \ldots, x_q\right)$ is equivalent to $Y = \left(y_1, y_2, \ldots, y_q\right)$ if $x_i = y_i$, $1 \leq i \leq q$, (denoted by $X =_m Y$). So, given the observed classification, say $Y^* = \left(y_1^*, y_2^*, \ldots, y_q^*\right)$, the p-value is obtained as:

$$p = \sum_{X \leq_m Y*} P_0[X = x] \tag{5}$$

where $P_0[X = x]$ denotes the probability mass function under the null hypothesis, $M(m, p_0)$. In Appendix A, a detailed calculation of the exact p-value is included.

This technique has been previously used for inference purposes involving multinomials. For instance, for the positional accuracy quality control of spatial data in [38,39] and in general for testing problems in contingency tables ([40,41], among others).

Once the p-value is calculated, the rule set is:

i. For a single column, the null hypothesis in Equation (3) should be rejected if the p-value obtained is less than the nominal value α.

ii. For L categories/columns, $2 \leq L \leq k$, we calculate the corresponding p-values for each testing problem. Assuming independence between the tests, we can state the global hypothesis of the quality requirements. To assure that the global size of the type I error is less than or equal to α, the Bonferroni criteria are used [42]. As a consequence, we should reject the global null hypothesis if at least one p-value is less than or equal to α/L.

As mentioned before, for each column of the QCCS, a different multinomial distribution can be proposed in the null hypothesis. This fact implies dealing with a set of multinomial distributions with equal or different amounts of classes (dimension) and equal or different levels in the quality requirement (probability vector). Binomial distributions are considered if all misclassification levels are merged to a single value.

4. Application

The data in Figure 2 and Table 1 are used to illustrate how this new structure QCCS works and how the exact tests are used for testing whether the levels of thematic accuracy are achieved or not. Figure 3 summarizes the specifications from Table 1 regarding the percentage of correctness and the limits on the misclassifications for each category. It presents the set of probability vectors, each one with values in the assumed order, and also groups some categories if the specifications do not analyze them separately. Figure 4 shows the same data as Figure 2, but the values in each column are ordered and grouped following Figure 3.

The required assumptions for these tests are as follows:

- The positional accuracy is adequate for the scale of the thematic product;
- The RDS is an adequate quality reference for this new approach of thematic accuracy quality control.

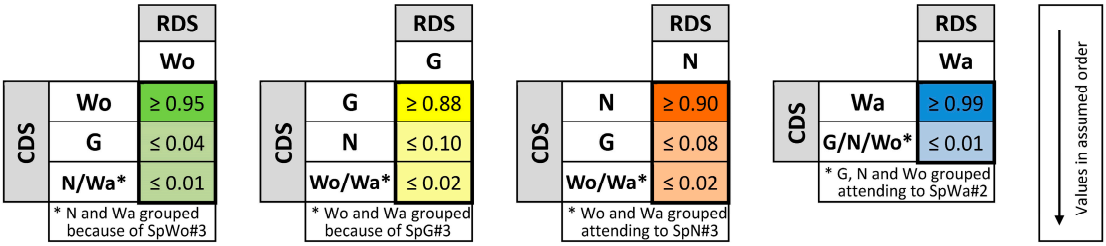

Figure 3. Quality specifications of Table 1 expressed as a set of probability vectors. Note that the values in each column are in the assumed freely eligible order. Also, note that some categories are grouped if the specifications do not analyze them separately. Wo = Woodland, G = Grassland, N = Non-vegetated, Wa = Water.

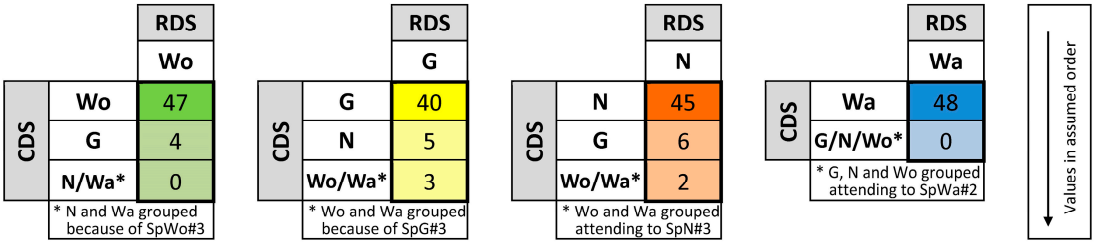

Figure 4. Quality control column set (QCCS) from Figure 2, prepared for the proposed exact multinomial test. In each column (i.e., vector), the values are reordered and grouped following Figure 3. Wo = Woodland, G = Grassland, N = Non-vegetated, Wa = Water.

Assuming the set of specifications presented in Table 1 (Figure 3), we define the corresponding null hypotheses for each column, and we illustrate the application of this new approach to this example.

a. Woodland.

According to Figure 3 (column Wo), for the *Woodland* category, we define a multinomial distribution with three categories: X_1 denoting the number of elements correctly classified; X_2, the number of elements confused with *Grassland*; and X_3, the number of elements classified as *Non-vegetated* or *Water*. Therefore, $q = 3$. The specifications lead us to the multinomial $M(51, p_0)$ with $p_0 = (0.95,\ 0.04,\ 0.01)$.

In this case, the observed classification is $Y^* = (47,\ 4,\ 0)$ (see Figure 4, column Wo), and the p-value is obtained by adding up the probabilities of $M(51, p_0)$ for classification outcomes that are worse according to the partial order $\leq_m$. The value obtained is 0.1678.

b. Grassland.

Along the same lines, from Figure 3 (column G), we define $X = (X_1, X_2, X_3)$, with X_1 being the number of elements correctly classified, X_2 being the number of misclassifications with *Non-vegetated*, and X_3 being the number of misclassifications with *Woodland* or *Water*. X follows $M(48, p_0)$ with a probability vector of $p_0 = (0.88,\ 0.10,\ 0.02)$.

The observed classification is $Y^* = (40, 5, 3)$ (see Figure 4, column G), and the p-value obtained is 0.2229.

c. Non-vegetated.

Following similar reasoning to that used in the previous categories, the multinomial distribution is $M(53, p_0)$ with $p_0 = (0.90,\ 0.08,\ 0.02)$, defined from the specifications shown in Figure 3 (column N).

The observed classification is $Y^* = (45,\ 6,\ 2)$ (see Figure 4, column N), and the p-value obtained is 0.1785.

d. Water.

In this category, as expressed in Table 1 and Figure 3 (column Wa), the maximum percentage of misclassified elements is 1%, so the classification results lead us to define a binomial distribution with the following parameters: the number of elements to be classified, $m = 48$, and the probability of correctness in the classification, $p = 0.99$, in short, $B(48, 0.99)$.

The observed classification is $Y^* = 48$ (see Figure 4, column Wa), and the p-value obtained is 1.

Overall decision. For this QCCS, four hypothesis tests were carried out, and in order to assure the same significance level, $\alpha = 0.05$, we applied the Bonferroni correction to compensate for the number of tests. As a consequence, we reject the hypothesis that the thematic quality levels specified in Table 1 and Figure 3 are globally achieved if any of the four p-values obtained are less than $\alpha/4 = 0.0125$. It does not occur in this example, and we can conclude that the data shown in Figure 4 globally satisfy the quality conditions stated in Table 1 (Figure 3). In addition, the individual analysis of each testing problem concludes that all of the quality requirements considered are accomplished.

In this example, we have illustrated how the QCCS can be applied and how it offers a more flexible and complete way to test statistically when a set of quality criteria are fulfilled, from both individual and global points of view. Such a detailed analysis would not be possible with the classical methods based on the confusion matrix.

5. Conclusions

In this study, a new approach for thematic accuracy quality control was presented. The approach is based on the assumption that the RDS does have the quality to be a reference, and this fact makes a more powerful and complete method for thematic accuracy quality control available than those based on a confusion matrix or on global indices. This method allows class by class quality control, which can consider not only the percentage of correctness but also the degree of misclassifications or confusion with other classes. It is a very flexible procedure because it allows us to simultaneously test the quality levels for a set of categories, and it also provides the possibility to merge classes when considering the misclassifications, which means the possibility of varying the dimensions of the underlying multinomial.

The proposed exact multinomial test for a category is based on a criterion for the achievement of the specifications. This criterion has been stated to be a partial order relation $X \leq_m Y$ between classification results X and specifications Y, which is similar to the lexicographic order. The procedure follows the same steps if the user changes the definition of $X \leq_m Y$ and, as a consequence, the calculation of the p-value of the corresponding testing hypothesis.

If a global hypothesis test is desired for L categories, the Bonferroni criteria are proposed. As a consequence, we should reject the global null hypothesis if at least one p-value is less than or equal to α/L.

The new approach can be widely applied to thematic accuracy quality control of any kind of spatial data in which the classification correctness should be assessed.

Author Contributions: Conceptualization, F.J.A.-L. and J.R.-A.; methodology, F.J.A.-L. and J.R.-A.; writing—original draft preparation, F.J.A.-L. and J.R.-A.; writing—review and editing, M.V.A.-F. and J.L.G.-B.; visualization, J.L.G.-B.; project administration, M.V.A.-F.

Funding: This work was supported by grant CMT2015-68276-R of the Spanish Ministry of Economy and Competitiveness.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

The calculation of the exact p-value considered in Section 3 is shown here. For the sake of simplicity, we consider an appealing case that is easy to perform, but the general situation follows the same steps.

Let us consider that we are interested in testing whether the classification proportions of corrections and proportions of misclassification agree with the following quality statements, ordered by importance by the user:

1. The proportion of correctness in class A is greater than p_{10}, $0 \leq p_{10} \leq 1$;
2. The percentage of misclassifications between classes A and B is less than p_{20}, $0 \leq p_{20} \leq 1$, $p_{10} \geq p_{20}$;
3. The percentage of misclassifications between classes A and C is less than p_{30}, $0 \leq p_{30} \leq 1$, $p_{10} \geq p_{20} \geq p_{30}$.

These quality control requirements can be stated as the following null hypothesis:

$\mathbb{H}_0$: The classification results agree with the set of specifications.

Let us suppose we have classified $m = 10$ elements, and the probability vector associated with the statements is $p_0 = (0.8, 0.1, 0.1)$. In the set of possible outcomes $\left\{0 \leq x_i \leq 10,\ i = 1,2,3 : \sum_{i=1}^{3} x_i = 10\right\}$, a result $X = (x_1, x_2, x_3)$ is worse than $Y = (y_1, y_2, y_3)$, $(X \leq_m Y)$ if

$(x_1 < y_1) \vee \{(x_1 = y_1) \wedge (x_2 > y_2)\} \vee \{(x_1 = y_1) \wedge (x_2 = y_2) \wedge (x_3 > y_3)\}$.

So, given the observed classification results, say $Y^* = (6,2,2)$, the p-value is obtained as

$$p = \sum_{X \leq_m Y^*} P_0[X = x], \tag{A1}$$

where $P_0[X = x]$ denotes the probability mass function under the null hypothesis, $M(10, p_0)$, with $p_0 = (0.8, 0.1, 0.1)$. The cases included in the sum, including the observed ones (6,2,2), are displayed in Table A1. The p-value is 0.1539, so we cannot reject the null hypothesis.

Table A1. Calculation of the exact p-value.

Outcomes	Condition	$P_0[X=x]$	Accumulated Probability
(6,2,2)	Outcome observed	0.0330301	0.0330301
(5,5,0)	$x_1 < 6$	0.0008257	0.0613809
(5,4,1)		0.0041287	0.0655096
(5,3,2)		0.0082575	0.0737671
(5,2,3)		0.0082575	0.0820246
(5,1,4)		0.0041287	0.0861533
(5,0,5)		0.0008257	0.0869790
(4,6,0)		0.0000860	0.0870650
...		...	...
(6,3,1)	$x_1 = 6,\ x_2 > 2$	0.0220201	0.0550502
(6,4,0)		0.0055050	0.0605552
	p-value		0.1539

References

1. OXERA. *What is the Economic Impact of Geoservices? Prepared for Google*; Oxera Consulting Ltd.: Oxford, UK, 2013; Available online: https://www.oxera.com/wp-content/uploads/2018/03/What-is-the-economic-impact-of-Geo-services_1-1.pdf (accessed on 24 July 2019).
2. Daly, C. Guidelines for assessing the suitability of spatial climate data sets. *Int. J. Climatol.* **2006**, *26*, 707–721. [CrossRef]
3. Devillers, R.; Stein, A.; Bédard, Y.; Chrisman, N.; Fisher, P.; Shi, W. Thirty Years of Research on Spatial Data Quality: Achievements, Failures, and Opportunities. *Trans. GIS* **2010**, *14*, 387–400. [CrossRef]
4. ISO. *ISO 19113:2002 Geographic Information—Quality Principles*; International Organization for Standardization: Geneva, Switzerland, 2002.
5. ISO. *ISO 5725-1:1994 Accuracy (Trueness and Precision) of Measurement Methods and Results—Part 1: General Principles and Definitions*; International Organization for Standardization: Geneva, Switzerland, 1994.

6. ISO. *ISO 19157:2013 Geographic Information—Data Quality*; International Organization for Standardization: Geneva, Switzerland, 2013.
7. Meidinger, D.V. *Protocol for Accuracy Assessment of Ecosystem Maps, British Columbia Ministry of Forest, Forest Science Program*; Technical Report 011; Crown Publications: Victoria, BC, USA, 2003; Available online: http://www.for.gov.bc.ca/hfd/pubs/Docs/Tr/Tr011.htm (accessed on 24 July 2019).
8. Olofsson, P.; Foody, G.M.; Herold, M.; Stehman, S.V.; Woodcock, C.E.; Wulder, M.A. Good practices for estimating area and assessing accuracy of land change. *Remote Sens. Environ.* **2014**, *148*, 42–57. [CrossRef]
9. Lea, C.; Curtis, A.C. *Thematic Accuracy Assessment Procedures. National Park Service Vegetation Inventory*; Version 2.0; Natural Resource Report NPS/NRPC/NRR—2010/204; US Department of the Interior, National Park Service: Fort Collins, CO, USA, 2010. Available online: https://www1.usgs.gov/vip/standards/NPSVI_Accuracy_Assessment_Guidelines_ver2.pdf (accessed on 24 July 2019).
10. Congalton, R.G.; Green, K. *Assessing the Accuracy of Remotely Sensed Data: Principles and Practices*, 2nd ed.; Lewis Publishers: Boca Raton, FL, USA, 2009.
11. Stehman, S.V.; Czaplewski, R. Design and Analysis for Thematic Map Accuracy Assessment: Fundamental Principles. *Remote Sens. Environ.* **1998**, *64*, 331–344. [CrossRef]
12. Wulder, M.A.; Franklin, S.E.; White, J.C.; Linke, J.; Magnussen, S. An accuracy assessment framework for large-area land cover classification products derived from medium-resolution satellite data. *Int. J. Remote Sens.* **2006**, *27*, 663–683. [CrossRef]
13. Strahler, A.H.; Boschetti, L.; Foody, G.M.; Friedl, M.A.; Hansen, M.C.; Herold, M.; Mayaux, P.; Morisette, J.T.; Stehman, S.V.; Woodcock, C.E. *Global Land Cover Validation: Recommendations for Evaluation and Accuracy Assessment of Global Land Cover Maps*; GOFC-GOLD Report No. 25; European Commision, Directorate-General Joint Research Centre, Institute for Environment and Sustainability: Ispra, Italy, 2006; Available online: https://publications.europa.eu/s/kVoK (accessed on 24 July 2019).
14. ISO. *ISO 19131:2007 Geographic Information—Data Product Specifications*; International Organization for Standardization: Geneva, Switzerland, 2007.
15. Ariza-López, F.J. *Fundamentos de Evaluación de la Calidad de la Información Geográfica*; Universidad de Jaén: Jaén, Spain, 2013.
16. Congalton, R.G. Accuracy assessment of remotely sensed data: Future needs and directions. In Proceedings of the Pecora 12 Symposium: Land Information from Space-Based Systems, Sioux Falls, South Dakota, 24–26 August 1993; American Society for Photogrammetry and Remote Sensing: Bethesda, MD, USA, 1994; pp. 383–388.
17. Foody, G.M. Status of land cover classification accuracy assessment. *Remote Sens. Environ* **2002**, *80*, 185–201. [CrossRef]
18. Comber, A.; Fisher, P.; Brunsdon, C.; Khmag, A. Spatial analysis of remote sensing image classification accuracy. *Remote Sens. Environ.* **2012**, *127*, 237–246. [CrossRef]
19. Liu, C.; Frazier, P.; Kumar, L. Comparative assessment of the measures of thematic classification accuracy. *Remote Sens. Environ.* **2007**, *107*, 606–616. [CrossRef]
20. Congalton, R.; Green, K. A practical look at the sources of confusion in error matrix generation. *Photogramm. Eng. Remote Sens.* **1993**, *59*, 641–644.
21. Smits, P.C.; Dellepiane, S.G.; Schowengerdt, R.A. Quality assessment of image classification algorithms for land-cover mapping: A review and proposal for a cost-based approach. *Int. J. Remote Sens.* **1999**, *20*, 1461–1486. [CrossRef]
22. Bird, A.C.; Taylor, J.C.; Brewer, T.R. Mapping National Park landscape from ground, air and space. *Int. J. Remote Sens.* **2000**, *21*, 2719–2736. [CrossRef]
23. Khorram, S. *Accuracy Assessment of Remote Sensing-Derived Change Detection*; American Society for Photogrammetry and Remote Sensing: Bethesda, MD, USA, 1999.
24. Abrams, M.; Bianchi, R.; Pieri, D. Revised mapping of lava flows on Mount Etna, Sicily. *Photogramm. Eng. Remote Sens.* **1996**, *62*, 1353–1359.
25. Bauer, M.E.; Burk, T.E.; Ek, A.R.; Coppin, P.R.; Lime, S.D.; Walsh, T.A.; Walters, D.K. Satellite inventory of Minnesota forest resources. *Photogramm. Eng. Remote Sens.* **1994**, *60*, 287–298.
26. Bowers, T.L.; Rowan, L.C. Remote mineralogic and lithologic mapping of the Ice River Alkaline Complex, British Columbia, Canada using AVIRIS data. *Photogramm. Eng. Remote Sens.* **1996**, *62*, 1379–1385.

27. Merchant, J.W.; Yang, L.; Yang, W. Validation of continental scale land cover data bases developed from AVHRR data. In Proceedings of the Pecora 12 Symposium: Land Information From Space-Based Systems, Sioux Falls, South Dakota, 24–26 August 1993; American Society for Photogrammetry and Remote Sensing: Bethesda, MD, USA, 1994; pp. 63–72.
28. Thierry, B.; Lowell, K. An uncertainty-based method of photo-interpretation. *Photogramm. Eng. Remote Sens.* **2001**, *67*, 65–72.
29. Congalton, R.G.; Gu, J.; Yadav, K.; Thenkabail, P.; Ozdogan, M. Global Land Cover Mapping: A Review and Uncertainty Analysis. *Remote Sens.* **2014**, *6*, 12070–12093. [CrossRef]
30. Ban, Y.; Gong, P.; Giri, C. Global land cover mapping using Earth observation satellite data: Recent progresses and challenges. *ISPRS J. Photogramm. Remote Sens.* **2015**, *103*, 1–6. [CrossRef]
31. Glick, H.B.; Routh, D.; Bettigole, C.; Seegmiller, L.; Kuhn, C.; Oliver, C.D. Modeling the Effects of Horizontal Positional Error on Classification Accuracy Statistics. *Photogramm. Eng. Remote Sens.* **2016**, *82*, 789–802. [CrossRef]
32. ASPRS. ASPRS Positional Accuracy Standards for Digital Geospatial Data. *Photogramm. Eng. Remote Sens.* **2015**, *81*, A1–A26. [CrossRef]
33. McRoberts, R.E.; Stehman, S.V.; Liknes, G.C.; Næsset, E.; Sannier, C.; Walters, B.F. The effects of imperfect reference data on remote sensing-assisted estimators of land cover class proportions. *ISPRS J. Photogramm. Remote Sens.* **2018**, *142*, 292–300. [CrossRef]
34. Ariza-López, F.J. *Evaluación de la Calidad de Diversas Series Cartográficas del Instituto de Cartografía de Andalucía: Informe Correspondiente a la Fase de Definición de Métodos*; Instituto de Cartografía de Andalucía: Sevilla, Spain, 2004.
35. Defourny, P.; Bontemps, S.; Obsomer, V.; Schouten, L.; Bartalev, S.; Herold, M.; Bicheron, P.; van Bogaert, E.; Leroy, M.; Arino, O. Accuracy assessment of global land cover maps: Lessons learnt from the GlobCover and GlobCorine experiences. In Proceedings of the 2010 European Space Agency Living Planet Symposium, Bergen, Norway, 28 June–2 July 2010; ESA Communication Production Office: Noordwijk, The Netherlands, 2011.
36. Yang, Y.; Xiao, P.; Feng, X.; Li, H. Accuracy assessment of seven global land cover datasets over China. *ISPRS J. Photogramm. Remote Sens.* **2017**, *125*, 156–173. [CrossRef]
37. Senseman, G.M.; Bagley, C.F.; Tweddale, S.A. *Accuracy Assessment of the Discrete Classification of Remotely-Sensed Digital Data for Landcover Mapping*; USACERL Technical Report EN-95/04; US Army Corps of Engineers, Construction Engineering Research Laboratories: Champaign, IL, USA, 1995; Available online: http://www.dtic.mil/get-tr-doc/pdf?AD=ADA296212 (accessed on 24 July 2019).
38. Ariza-López, F.J.; Rodríguez-Avi, J.; Alba-Fernández, V. Ariza-López, F.J.; Rodríguez-Avi, J.; Alba-Fernández, V. A Positional Quality Control Test Based on Proportions. In *Geospatial Technologies for All, AGILE 2018, Lecture Notes in Geoinformation and Cartography*; Mansourian, A., Pilesjö, P., Harrie, L., van Lammeren, R., Eds.; Springer: Cham, Switzerland, 2018; Rodríguez-Avi, J. [CrossRef]
39. Ariza-López, F.J.; Rodríguez-Avi, J.; González-Aguilera, D.; Rodríguez-Gonzálvez, P. A New Method for Positional Accuracy control for Non-Normal Errors Applied to Airborne Laser Scanner Data. *App. Sci.* **2019**, *9*, 3887. [CrossRef]
40. Mehta, C.R.; Patel, N.R. A Network Algorithm for Performing Fisher's Exact Test in r × c Contingency Tables. *J. Am. Stat. Assoc.* **1983**, *78*, 427–434. [CrossRef]
41. Storer, B.E.; Choongrak, K. Exact properties of some exact test statistics for comparing two binomial proportions. *J. Am. Stat. Assoc.* **1990**, *85*, 146–155. [CrossRef]
42. Goeman, J.J.; Solari, A. Multiple hypotheses testing in genomics. *Stat. Med.* **2014**, *33*, 1946–1978. [CrossRef] [PubMed]