Scientific Research Publishing

Journal of Geographic Information System, 2020, 12, 178-187
https://www.scirp.org/journal/jgis
ISSN Online: 2151-1969
ISSN Print: 2151-1950

On Statistical Measures for Data Quality Evaluation

Xiaoxia Han

Department of Public Health Sciences, Henry Ford Health System, Detroit, USA
Email: xhan2@hfhs.org

How to cite this paper: Han, X. (2020) On Statistical Measures for Data Quality Evaluation. *Journal of Geographic Information System*, **12**, 178-187.
https://doi.org/10.4236/jgis.2020.123011

Received: April 28, 2020
Accepted: June 8, 2020
Published: June 11, 2020

Abstract

Most GIS databases contain data errors. The quality of the data sources such as traditional paper maps or more recent remote sensing data determines spatial data quality. In the past several decades, different statistical measures have been developed to evaluate data quality for different types of data, such as nominal categorical data, ordinal categorical data and numerical data. Although these methods were originally proposed for medical research or psychological research, they have been widely used to evaluate spatial data quality. In this paper, we first review statistical methods for evaluating data quality, discuss under what conditions we should use them and how to interpret the results, followed by a brief discussion of statistical software and packages that can be used to compute these data quality measures.

Keywords

GIS Data Quality, Sensitivity, Specificity, Kappa, Weighted Kappa, Bland-Altman Analysis, Intra-Class Correlation Coefficient

1. Introduction

Spatial data quality is limited by the quality of the data sources such as traditional paper maps or more recent remote sensing data [1]. Spatial operations and spatial analyses such as projection, overlay, buffering, network analysis, and spatial regression, highly depend on the quality of spatial data. Without a prior knowledge of data quality, it is difficult to conduct downstream operations and thus make it hard to make informed decisions. Therefore, data quality is an important aspect for geographical information systems (GIS) databases, and has drawn considerable attention from academic communities, government agencies as well as industry [2].

There are four levels of measurement scales that are used to capture spatial

data: nominal, ordinal, interval and ratio. Normal and ordinal data belong to categorical data, while interval and ratio data belong to numerical data. In the past several decades, different statistical measures have been developed to evaluate data quality for different types of data. Although these methods were originally developed for medical research or psychological research [3]-[8], they have been widely used to evaluate spatial data quality [9] [10] [11] [12]. In this paper, we first review these different statistical methods for evaluating data quality for different types of spatial data, discuss under what conditions we should use them and how to interpret the results, followed by a brief discussion of statistical software and packages that can be used to compute these data quality measures.

2. Methods Used to Measure Data Quality

2.1. Nominal Categorical Data

Nominal categorical data is used to label variables without providing any quantitative value, which is the simplest form of a scale of measure. Unlike ordinal data, nominal data cannot be ordered. For example, land cover/land use can be categorized into "open water", "residential", "commercial", "wetland", "mixed forest", "agriculture", and there is no inherent order among these categories. Without loss of generality, we first consider a simple classification problem where there are only two categories. For example, we have a map of a certain mineral and we want to evaluate the accuracy of the mineral map. The data can be summarized in a 2-by-2 confusion or error matrix that cross-tabulates the truth and classification on the map (Table 1). A variety of accuracy measures can be derived from a 2-by-2 confusion matrix [3]. Table 2 gives a list of these

Table 1. A confusion matrix for 2-class classification problem.

		Truth	
		Mineral present	**Mineral absent**
Classification on the map	Mineral present	a	b
	Mineral absent	c	d

Table 2. Accuracy measures for 2-by-2 confusion matrix.

Accuracy measures	Interpretation	How to calculate
True positive (TP)	Mineral correctly identified as present on the map	a
False positive (FP)	Non-mineral incorrectly identified as present on the map	b
True negative (TN)	Non-mineral correctly identified as absent on the map	d
False negative (FN)	Mineral incorrectly identified as absent on the map	c
Sensitivity	Conditional probability that a true "present" is correctly classified on the map	TP/(TP + FN) = a/(a + b)
Specificity	Conditional probability that true "absent" is correctly classified on the map	TN/(TN + FP) = d/(b + d)

accuracy measures.

For multi-class classification, we can use one against all approach for TP, TN, FP, FN. Suppose we have a map of classification of the likelihood of landslides as shown in Table 3. There are three classes: low, moderate, and high. TP of low is all low instances that are classified and low on the map. TN of low is all non-low (*i.e.*, moderate and high) instances that are not classified as low. FP of low is all non-low instances that are classified as low, and FN of low is all low instances that are not classified as low. Similarly, we can calculate TP, FN, TN, FP for moderate and high categories, respectively. Sensitivity and specificity can also be calculated based on TF, FN, TN, and FP. Correct classification rate, misclassification rate can also be calculated for confusion matrix with two or more categories.

Correct classification rate is the number of correct classified instances on the map divided by the total number of instances, *i.e.*, the sum of number on the diagonal divided by N, where N is the total number of instances. Misclassification rate is the number of incorrect classified instances on the map divided by the total number of instances, *i.e.*, the sum of number off-diagonal divided by total instance N.

Kappa index can be used to evaluate attribute accuracy when truth is known [4]. Intuitively, Kappa index represents the truth and map agreement taking into account the expected agreement by chance. We assume to have a k-by-k confusion matrix M, and create a proportions matrix P, which is M/n. Let $p_{i,j}$ be the proportion of observations in row *i*, column *j*, p_{i+} be the proportion of mapped data in row (class) *i*, and p_{+j} be the proportion of mapped data in column j. We further define $p_o = \sum_{i=1}^{k} p_{ii}$, and $p_c = \sum_{i=1}^{k} p_{i+} p_{+j}$. Then Kappa index can be calculated as $\hat{K} = (p_o - p_c)/(1 - p_c)$. The Kappa index can take values from −1 to 1. The interpretation is somewhat arbitrary (Table 4). Negative values indicate that the observed agreement is worse than what would be expected by change alone.

When the truth data is not available, Kappa index can be used to evaluate relative agreement between two data sources, or pairwise relative agreement among more than two data sources. If Kappa index is small between two data sources, we can infer that the data quality of at least of one data source is not good. If we have 3 data sources, two of them have "good kappa", but both of them have "bad kappa" with the third data sources, we can infer that the first two data sources has similar data quality – either both of them are good or both of them are bad. In this case, other information needs to be collected to determine quality for

Table 3. A confusion matrix for 3-class classification problem.

		Truth		
		Low	**Moderate**	**High**
Classification on the map	Low	70	10	5
	Moderate	8	67	20
	High	1	10	14

Table 4. Interpretation of Kappa index.

Kappa index value	Interpretation
0	Agreement equivalent to chance
0.10 - 0.20	Slight agreement
0.21 - 0.40	Fair agreement
0.41 - 0.60	Moderate agreement
0.61 - 0.80	Substantial agreement
0.81 - 0.99	Near-perfect agreement
1.00	Perfect agreement

these three data sources.

2.2. Ordinal Categorical Data

Ordinal data is a categorical data type that does not have a number (*i.e.*, not quantitative), but the data have natural, ordered categories. For example, average temperature can be classified as "very cold", "cold", "chilly", "lukewarm ", "warm", "hot", "very hot" on a map, or landslides incidence of a certain area can be shown on a map with different color to indicate "low", "moderate" and "high" likelihood of landslides. In other words, although ordinal data do not represent a quantity, but they do have an inherent order.

The Kappa index we discussed previously is not appropriate for ordinal categorical data, because it assumes all the errors in the confusion matrix is considered of equal importance. However, for ordinal data, the classification errors vary in their importance. In other words, the "costs" of misclassification are different among the ordinal categorical data. For example, it may be far worse to classify an area with high likelihood of landslides area to low likelihood of landslides than to classify it as a moderate likelihood of landsides. In this scenario, the weighted Kappa is the correct index to use for evaluating data quality purpose [5]. The following describes the procedure to calculate a weighted Kappa index with the confusion matrix in Table 3.

To calculate weight Kappa, we need to create another Weights matrix which contains the weights for each cell. The diagonal cell in the Weights matrix is 1, indicating full credit for each class correctly. The value of off-diagonal cells should be assigned values by the analyst, with weight value between 0 and 1. A value of 0 means that there is no partial credit for misclassification for one class to the other, a value of 1 means we give full credit for misclassification (*i.e.*, we treat this misclassification as correct classification). Any value less than 1 but greater than 0 means there is partial credit for misclassification.

Table 5 gives a hypothetical 3-by-3 Weights matrix for landslides likelihood classification. In this example, we give full credit for correct classification, as shown that all the diagonal elements are 1. We give partial credit (0.5) for classifying "low" as "moderate", and partial credit (0.2) for classifying "low" as "high".

Table 5. Weights matrix for landslides likelihood classification.

		Truth		
		Low	**Moderate**	**High**
Classification on the map	Low	1	0	0
	Moderate	0.5	1	0
	High	0.2	0.5	1

We also give partial credit for classifying "moderate" to "high". However, we do not give partial credit for classifying "moderate" to "low" or misclassifying "high". In general, the weighted Kappa can be calculated as following.

We assume to have a k-by-k confusion matrix M, and create a proportions matrix P, which is *M*/*n*. Let $p_{i,j}$ be the proportion of observations in row *i*, column *j*, p_{i+} be the proportion of mapped data in row (class) *i*, and p_{+j} be the proportion of mapped data in column *j*. Let w_{ij} denote the weight assigned to the *i*,*j* th element in matrix W. We further define $p_0^* = \sum_{i=1}^{k}\sum_{j=1}^{k} w_{ij} p_{ij}$, and $p_c^* = \sum_{i=1}^{k}\sum_{j=1}^{k} w_{ij} p_{i+} p_{+j}$. Then the weighted Kappa can be defined as $\hat{K}_w = \left(p_0^* - p_c^*\right) \Big/ \left(1 - p_c^*\right)$.

2.3. Numerical Data

Numerical data or quantitative data is a numerical measurement that can be represented in numbers. Numerical data can be discrete or continuous. Discrete data represent times that can be counted, and it has a finite number of possible values and the values cannot be subdivided meaningfully. For example, the number of people in a census tract is discrete numerical data, and the number of houses in a certain area is also discrete numerical data. On the other hand, continuous data represent measurement that can be meaningfully subdivided into finer and finer increments, depending upon the precision of the measurement system. For example, the annual precipitations and temperature are both continuous data. Bland-Altman analysis [6] and intra-class correlation coefficient (ICC) [7] [8] are the two widely used methods to assessing agreement between measures of numerical data.

Bland-Altman plot is a scatter plot of the difference between two measurements (Y-axis) against the average of two measurements (X-axis), with 95% limits of agreement. The limits of agreement are calculated by the mean observed difference ± 1.96 X standard deviation of observed difference. Consider a situation where we developed a new algorithm to process images, which is computationally more efficient than the standard method. We want to assess the agreement between intensity values from this new image processing algorithm (observed value) and the ground truth from standard method. The true values with sample size n = 30 were simulated from uniform distribution (0, 255) and the observed values were obtained by the true values plus values that simulated from a normal distribution with sample size n = 30, mean 0 and standard deviation 3.

The 30 pairs of true and observed values are shown in Table 6.

Figure 1 shows the Bland-Altman plot using the data in Table 6. The X-axis is the average of true and observed values. The Y-axis is the difference between true and observed values. The mean of observed difference is −0.2, which is shown as the dash line just below the solid line with difference being 0. The standard deviation of observed difference is 3.36. The limits of agreements are (−6.79, 6.39), which are represented by the other two dash lines far from the solid line. Note that there is one data point outside the 95% limits of agreement. This plot implied that the intensity data from the new algorithm could vary from the true values by −6.79 to 6.39 for the 95% of the data points. For the 5% of the

Table 6. An artificial example of image signal intensity obtained by new algorithm and ground truth values using simulation data.

Sample#	Truth	Observed	Sample#	Truth	Observed	Sample#	Truth	Observed
1	53	53	11	41	41	21	187	186
2	74	74	12	35	35	22	42	41
3	115	118	13	137	133	23	130	130
4	182	184	14	77	76	24	25	28
5	40	42	15	154	155	25	53	55
6	180	183	16	100	104	26	77	77
7	189	191	17	144	144	27	3	2
8	132	132	18	198	199	28	76	78
9	126	120	19	76	76	29	174	176
10	12	14	20	155	151	30	68	66

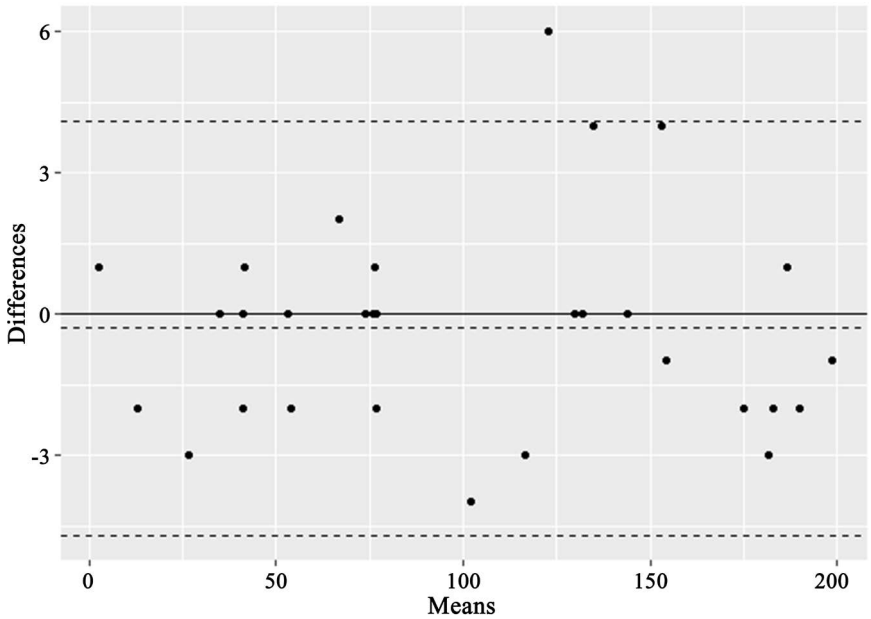

Figure 1. Bland-Altman plot for comparing data from new algorithm to the ground truth using simulation data from Table 6.

data points, the variations could be outside these limits. It seems like the new algorithm cannot be used to substitute the standard method. Note that there is no uniform criterion on acceptable values of limits of agreement. This depends on the variables being measured and researchers should use their domain knowledge to make decisions.

Intra-class correlation coefficient (ICC) is a widely used index to assess agreement between two numerical measures. ICC provides an estimate of overall concordance between data from two or more sources. It is somewhat akin to "analysis of variance". There are 10 forms of ICCs, depending on the model selection (random effect model vs. mixed effect model) and type of selection (single measurement or multiple measurements), and definition of selection (absolute agreement or consistency). A comprehensive review of selecting and reporting ICCs can be found in Koo and Li [13].

Importantly, we need to know that there are no standard values for acceptable reliability based on ICC. A low ICC may due to lack of variability among the sampled data, instead of low degree of agreement between two methods or two raters. Thus, it is suggested to have at least 30 samples when using ICC to evaluate agreement. The interpretation of ICC values is somewhat arbitrary (Table 7).

Using the simulated data in Table 6, the ICC based on mixed effect model with absolute agreement and single measurement, the ICC is 0.999. According to Table 6, this ICC could be interpreted as excellent agreement. This is somewhat contradicted to what we found using Bland-Altman analysis. Bland-Altman provides more information for decision making, as it gives us a more comprehensive summary of the data (mean of the difference between the two methods, standard deviation of the difference, 95% limits of agreements, etc) other than just one number to indicate the agreement measured by concordance using ICC. In addition, we have to understand that this ICC is only an expected value of true ICC based on the 30 data pairs in Table 6. It may be more interesting to conduct a hypothesis test to investigate whether the observed ICC value significantly exceeds some prespecified threshold.

When we do not have ground truth data, we can still use ICC to evaluate the agreement between two data sources. A high ICC means high agreement between the two data sources, they can have equally good data quality or equally bad data quality. A low ICC means low agreement between the two data sources, at least one of the data sources has bad data quality. Similarly, Bland-Altman Analysis can also be used to evaluating the agreement of two data sources/models

Table 7. Interpretation of intra-class correlation coefficient (ICC).

ICC	Interpretation
<0.5	Poor agreement
0.5 to <0.75	Moderate agreement
0.75 to <0.9	Good agreement
0.9 - 1.0	Excellent agreement

predictions when the ground truth is not available.

3. Statistical Software

Some of the statistical measures for data quality evaluation are relatively simple, and it is possible to calculate using traditional "pen and paper" approach. However, as the sample size increases, statistical software is needed to conduct such analysis. In addition, for the more complicated methods such as weighted Kappa index, Bland-Altman plot and ICC, we usually need statistical software to do the calculation. Statistical software such SAS software [14], SPSS [15], Stata [16], and R [17] can compute those statistical measures. However, only R is an open source software which means it is free to be used by anyone in any country in the world. R compiles and runs on a wide variety of UNIX platforms, Windows and MacOS. The package "EvaluationMeasures" [18] can calculate sensitivity, specificity, TP, FP, TN, FN, etc. The package "psych" [19] can calculate Kappa index, weighted Kappa and ICCs and conduct hypothesis testing. The packages "irr" [20] and "icc" [21] can also calculate different forms of ICCs. The package "blandr" [22] can carry out Bland-Altman analyses and produce plot.

4. Summary

Most GIS databases contain data errors. The data errors may come from human entry error, the source of data (paper maps or images), or imperfection of the image processing algorithms. These data imperfections have a direct impact on the reliability of spatial analysis results. For example, if objects have slightly different boundaries for a polygon overlay operation, a large number of "slivers" will be produced which will result in errors for downstream analysis [23]. Thus, spatial data quality has been identified as a critical issue for organizations.

Different GIS applications require different degree of details of spatial data which depends on the purpose of the applications. Importantly, we have to understand that there is no "one-size-fit-all" guideline for evaluating spatial data quality. Even if we use the same methods, we may use different "cut-off" values to decide whether the data quality has accuracy enough or not. It is important to choose the correct methods to evaluate data quality and wisely interpret the results, so that we can have a better knowledge of the data quality, which in turn, helps us to make informed decisions.

Conflicts of Interest

The author declares no conflicts of interest regarding the publication of this paper.

References

[1] Zeiler, M. (1999) Modeling Our World: The ESRI Guide to Geodatabase Design. ESRI, Inc., New York.

[2] Devillers, R., Stein, A., Bédard, Y., Chrisman, N., Fisher, P. and Shi, W. (2010)

Thirty Years of Research on Spatial Data Quality: Achievements, Failures, and Opportunities. *Transactions in GIS*, **14**, 387-400. https://doi.org/10.1111/j.1467-9671.2010.01212.x

[3] Fleiss, J.L., Levin, B. and Paik, M.C. (2013) Statistical Methods for Rates and Proportions. John Wiley & Sons, New York.

[4] Cohen, J. (1960) A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, **20**, 37-46. https://doi.org/10.1177/001316446002000104

[5] Cohen, J. (1968) Weighted Kappa: Nominal Scale Agreement Provision for Scaled Disagreement or Partial Credit. *Psychological Bulletin*, **70**, 213. https://doi.org/10.1037/h0026256

[6] Bland, J.M. and Altman, D. (1986) Statistical Methods for Assessing Agreement between Two Methods of Clinical Measurement. *The lancet*, **327**, 307-310. https://doi.org/10.1016/S0140-6736(86)90837-8

[7] Bartko, J.J. (1966) The Intraclass Correlation Coefficient as a Measure of Reliability. *Psychological Reports*, **19**, 3-11. https://doi.org/10.2466/pr0.1966.19.1.3

[8] Shrout, P.E. and Fleiss, J.L. (1979) Intraclass Correlations: Uses in Assessing Rater Reliability. *Psychological bulletin*, **86**, 420. https://doi.org/10.1037/0033-2909.86.2.420

[9] Fielding, A.H. and Bell, J.F. (1997) A Review of Methods for the Assessment of Prediction Errors in Conservation Presence/Absence Models. *Environmental Conservation*, **24**, 38-49. https://doi.org/10.1017/S0376892997000088

[10] Sinha, S. (2019) Bland-Altman Analysis for Evaluating AHP-Based Wildlife Habitat Suitability Models. *Research & Reviews*: *Journal of Space Science & Technology*, **4**, 11-18. https://doi.org/10.37591/.v4i2.1958

[11] Rhew, I.C., Vander Stoep, A., Kearney, A., Smith, N.L. and Dunbar, M.D. (2011) Validation of the Normalized Difference Vegetation Index as a Measure of Neighborhood Greenness. *Annals of Epidemiology*, **21**, 946-952. https://doi.org/10.1016/j.annepidem.2011.09.001

[12] Feng, X., Wang, Y., Chen, L., Fu, B. and Bai, G. (2010) Modeling Soil Erosion and Its Response to Land-Use Change in Hilly Catchments of the Chinese Loess Plateau. *Geomorphology*, **118**, 239-248. https://doi.org/10.1016/j.geomorph.2010.01.004

[13] Koo, T.K. and Li, M.Y. (2016) A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. *Journal of Chiropractic Medicine*, **15**, 155-163. https://doi.org/10.1016/j.jcm.2016.02.012

[14] SAS Institute (2015) Base SAS 9.4 Procedures Guide. SAS Institute.

[15] IBM Corp. Released (2017) IBM SPSS Statistics for Windows, Version 25.0. IBM Corp., Armonk, NY.

[16] StataCorp (2019) Stata Statistical Software: Release 16. StataCorp LLC, College Station, TX.

[17] Team, R.C. (2013) R: A Language and Environment for Statistical Computing.

[18] Khorsand, B., Zahiri, J. and Savadi, A. (2016) Evaluation Measures: Collection of Model Evaluation Measure Functions. https://cran.r-project.org/web/packages/EvaluationMeasures/index.html

[19] Revelle, W. and Revelle, M.W. (2015) Package 'Psych'. The Comprehensive R Archive Network.

[20] Gamer, M., Lemon, J., Gamer, M.M., Robinson, A. and Kendall's, W. (2012) Pack-

age 'IRR'. Various Coefficients of Interrater Reliability and Agreement.

[21] Wolak, M. and Wolak, M.M. (2015) Package "ICC". Facilitating Estimation of the Intraclass Correlation Coefficient.

[22] Datta, D. (2018) Blandr: Bland-Altman Method Comparison. https://cran.r-project.org/web/packages/blandr/index.html

[23] Bolstad, P. (2016) GIS Fundamentals: A First Text on Geographic Information Systems. Eider (Press Minnesota).