An Agreement Coefficient for Image Comparison

Lei Ji and Kevin Gallo

Abstract

Combination of datasets acquired from different sensor systems is necessary to construct a long time-series dataset for remotely sensed land-surface variables. Assessment of the agreement of the data derived from various sources is an important issue in understanding the data continuity through the time-series. Some traditional measures, including correlation coefficient, coefficient of determination, mean absolute error, and root mean square error, are not always optimal for evaluating the data agreement. For this reason, we developed a new agreement coefficient for comparing two different images. The agreement coefficient has the following properties: non-dimensional, bounded, symmetric, and distinguishable between systematic and unsystematic differences. The paper provides examples of agreement analyses for hypothetical data and actual remotely sensed data. The results demonstrate that the agreement coefficient does include the above properties, and therefore is a useful tool for image comparison.

Introduction

Since the launch of Landsat in 1972, the landmark in the use of remote sensing technology for earth observation, various satellite platforms and sensor systems have been developed to acquire spectral data from the land-surface. The spectral data received by the sensors have been used to detect and monitor land-surface characteristics and their changes over a large area. Over the past 30 years, tremendous land-surface data have been collected and archived through satellite sensors. The accumulation of the historical data records make possible analyzing inter- and intra-annual changes in land-surface characteristics throughout the world. To construct a relatively long time-series for spectral data, it is often necessary to combine datasets derived from different platforms and sensors. For example, the Landsat datasets including Multispectral Scanner (MSS), Thematic Mapper (TM) and Enhanced Thematic Mapper Plus (ETM+) were integrated to detect the U.S. land-cover changes since 1973 (Loveland *et al.*, 2002). Time series of the Advanced Very High Resolution Radiometer (AVHRR) data from NOAA-7, -9, -11, and -14 satellites were used to detect vegetation green-up trend from 1982 to 1999 in Eurasia and North America (Zhou *et al.*, 2001, 2003; Bogaert *et al.*, 2002). Data acquisition systems, however, differ in sensor design (spectral range, number of bands, bandwidth, radiometric resolution, ground resolution, and swath width), satellite platform characteristics (instrument, frequency of visit, and overpass time), and data processing technique (geometric and atmospheric corrections, geometric registration, resampling, and temporal compositing). Therefore, comparison of the datasets acquired from various sources becomes an important issue in understanding the data continuity through time-series. In the past a few years, special attention has been paid to the comparison of the normalized difference vegetation index (NDVI) derived from different satellite platforms and sensors (Goetz, 1997; Teillet *et al.*; 1997, 2001; Goward *et al.*, 2003). The comparisons of NDVI datasets have been including those between AVHRR and Moderate Resolution Imaging Spectroradiometer (MODIS) sensors. Some investigators compared NDVI data using the AVHRR and MODIS red and near-infrared bands simulated from close-range or airborne hyperspectral data (Gitelson and Kaufman, 1998; Gao, 2000; Steven *et al.*, 2003). Others made direct comparisons for AVHRR-NDVI and MODIS-NDVI images with the identical time interval and spatial resolution (Huete *et al.*, 2002; Gallo *et al.*, 2004).

To evaluate data agreement or disagreement as in the examples discussed above, some statistical methods were widely used, including Pearson correlation coefficient (r), coefficient of determination (r^2), mean absolute error (MAE), root mean square error (RMSE), and others. Those traditional measures, however, are not always optimal for evaluating the data agreement or disagreement (a brief review of the measures of agreement is given in the following section). For example, r or r^2 merely indicates the linear covariation between two datasets rather than the actual difference; MAE and RMSE are dimensional measures of disagreement, thus are not independent of data scale and unit. The measure of agreement developed by Willmott (1981, 1982) overcomes some disadvantages of the above-mentioned measures. But Willmott's measure is more appropriate for the investigation of model validation, where observed and model-predicted values need to be compared. In the latter situation, observed data are assumed error-free and treated as a reference of the comparison. For remotely sensed data, all datasets for comparison are assumed to contain some measurement errors. Thus, to develop a measure of agreement is necessary for the comparison among datasets derived from different sources. The objective of this study is to develop an "agreement

Lei Ji was formally with the Cooperative Institute for Research in the Atmosphere, Colorado State University and is currently with the Science Applications International Corporation, USGS Center for Earth Resources Observation and Science, Sioux Falls, SD 57198 (lji@usgs.gov).

Kevin Gallo is with the NOAA National Environmental Satellite, Data, and Information Service, Camp Springs, MD 20746 (kgallo@usgs.gov).

Photogrammetric Engineering & Remote Sensing
Vol. 72, No. 7, July 2006, pp. 823–833.

0099-1112/06/7207–0823/$3.00/0

coefficient," which is capable of quantifying actual difference between different datasets and separating systematic and unsystematic errors. The agreement coefficient should have the properties of non-dimensionality and boundedness, which are essential for cross-site data comparison. This paper introduces the concept and definition of the agreement coefficient and the relevant parameters, and provides examples of agreement analysis from hypothetical datasets and actual remotely sensed datasets.

A Review of Measures of Agreement

Several widely used measures of agreement and disagreement are briefly reviewed here. Details about r, r^2, MAE and RMSE can be found in most introductory statistics textbooks. Robinson's "coefficient of agreement," Willmott's "index of agreement," and Mielke's "measure of agreement" are introduced in detail in their publications (Robinson, 1957, 1959; Willmott, 1981, 1982; Willmott *et al.*, 1985; Mielke, 1984, 1991; Mielke *et al.*, 1996, 1997).

Correlation Coefficient and Coefficient of Determination

Pearson correlation coefficient (r) and coefficient of determination (r^2) are reported most frequently in literature to indicate agreement between different datasets. Correlation coefficient measures linear covariation between two datasets. Thus, higher r or r^2 indicates that the two datasets have similar spatial or temporal patterns. Usually, r^2 provides more useful expression of information than r, because r^2 measures proportion of the total data variation explained by the regression model. Both r and r^2 have the advantages that they are non-dimensional and bounded (r is bounded by -1 and 1, and r^2 is bounded by 0 and 1). Non-dimensionality and boundedness are advantageous when r and r^2 values are compared across datasets with different units or scales. But, r and r^2 may be misleading when they are used to measure data agreement, because they fail to measure actual difference between two datasets. For example, r and r^2 for datasets A (5, 12, 7, 5, 4, 15, 11, 9) and B (225, 505, 305, 225, 185, 625, 465, 385) are 1, showing a perfect linear covariation ($B = 40A + 25$). But the deviation between A and B is very high, indicated by the difference between A and B. Thus, neither r nor r^2 is considered an appropriate measure for data agreement.

Mean Bias Error, Mean Absolute Error, and Root Mean Square Error

The mean bias error (MBE) is defined as

$$MBE = \frac{1}{n}\sum_{i=1}^{n}(X_i - Y_i), \tag{1}$$

where X_i and Y_i are the i^{th} observations of datasets X and Y, and n is number of the observations. MBE measures the average difference between the two datasets. But, as an average value, it eliminates the positive and negative differences between observations. In fact, it is merely the difference between the mean values of the two datasets:

$$MBE = \frac{1}{n}\sum_{i=1}^{n}(X_i - Y_i) = \overline{X} - \overline{Y}, \tag{2}$$

where $\overline{X}$ and $\overline{Y}$ are the mean values of datasets X and Y. Another two more meaningful indices are mean absolute error (MAE) and root mean square error (RMSE), which are given by

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|X_i - Y_i|, \tag{3}$$

and

$$RMSE = \left[\frac{1}{n}\sum_{i=1}^{n}(X_i - Y_i)^2\right]^{\frac{1}{2}}; \tag{4}$$

MBE, MAE and RMSE measure actual difference between two different datasets. When datasets X and Y are in perfect agreement, or all X_i's and Y_i's are identical, MBE, MAE and RMSE equal zero. These measures, however, are not standardized and not bounded. Moreover, they are dependent on data unit and scale, which can make cross-site comparison difficult. For example, if datasets X and Y are recorded in 8-bit bytes (values range from 0 to 255), or in 16-bit bytes (0 to 65535), the values of MBE, MAE, and RMSE between X and Y are different.

Mean Absolute Percent Error and Root Mean Squared Percentage Error

The mean absolute percent error (MAPE) and root mean squared percentage error (RMSPE) are measures used to validate forecast models (Ramanathan, 1995). These measures are defined as

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}100\,\frac{|X_i - Y_i|}{X_i}, \tag{5}$$

and

$$RMSPE = \left[\frac{1}{n}\sum_{i=1}^{n}\left(100\,\frac{X_i - Y_i}{X_i}\right)^2\right]^{\frac{1}{2}}. \tag{6}$$

The difference between MAPE and MAE and between RMSPE and RMSE is that MAPE and RMSPE are standardized values, and are independent of the unit of the measurement. But, MAPE is meaningful only if all X_i values are positive (Ramanathan, 1995). Moreover, MAPE and RMSPE are very unstable when X_i values are near zero. Negative and near-zero values are common in remotely sensed data, especially for some indices (e.g., NDVI) calculated from the differences or ratios of the reflectance in different bands. Finally, the MAPE and RMSPE are not symmetric metrics. If X and Y are interchanged in calculation, MAPE and RMSPE generate different values.

Willmott's Index of Agreement

To circumvent the problems associated with r, r^2, MAE, and RMSE, Willmott (1981, 1982) developed the index of agreement, which is used especially for validating prediction models. The index of agreement (d) is expressed as

$$d = 1 - \frac{\sum_{i=1}^{n}(X_i - Y_i)^2}{\sum_{i=1}^{n}\left(|X_i - \overline{X}| + |Y_i - \overline{X}|\right)^2}, \tag{7}$$

where X_i is the observed value, Y_i is the modeled or simulated value, $\overline{X}$ is the mean of observed values. The numerator of the main term in Equation 7, $\Sigma(X_i - Y_i)^2$, is sum of square error (SSE). The denominator, sum of the squared absolute distances from X_i to $\overline{X}$ and from Y_i to $\overline{X}$, is referred to as potential error (PE). In fact, PE is dependent on the range of X and Y, and used to standardized the SSE. Thus, d is a bounded and non-dimensional measure. If all modeled values fit the observed values, d equals 1. Mean square error (MSE) is the arithmetic mean of SSE. Willmott (1981, 1982) also defined two other error indices, the systematic MSE (MSE_s)

$$MSE_s = \frac{1}{n}\sum_{i=1}^{n}(X_i - \hat{Y}_i)^2, \quad (8)$$

and the unsystematic MSE (MSE_u)

$$MSE_u = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2, \quad (9)$$

where $\hat{Y}$ is the predicted Y value obtained from regression model $\hat{Y} = a + bX$. The MSE_s refers to the systematic error that can be predicted and thus adjusted by refining the model, whereas the MSE_u refers to the random error that cannot be predicted or adjusted. The sum of MSE_s and MSE_u is MSE:

$$MSE = MSE_s + MSE_u. \quad (10)$$

The proportions of systematic and unsystematic errors to the total errors can be derived from MSE_s/MSE and MSE_u/MSE. The attractiveness of Willmott's indices is the partition of total error into systematic and unsystematic errors. But Willmott's indices are not symmetric as indicated by Equation 7. Interchanging X and Y can generate different d values, because mean values of X and Y are usually different.

Mielke's Measure of Agreement

Mielke (1984, 1991) developed the measure of agreement, a method to validate prediction models. The measure of agreement (ρ) is given by

$$\rho = 1 - \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - Y_i)^2}{\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}(X_i - Y_j)^2}. \quad (11)$$

The numerator of the main term in Equation 11 is MSE. The denominator represents the average value of MSE over all $n!$ equally likely permutation of $X_1, \ldots, X_n$ relative to $Y_1, \ldots, Y_n$ under the null hypothesis that the n pairs (X_i and Y_i for $i = 1, \ldots, n$) are merely the result of random assignment (Mielke, 1991; Mielke *et al.*, 1996). Mielke's ρ is dimensionless, and bounded above by 1. Note that ρ is symmetric, and thus it produces an identical value if the positions of X and Y are switched.

Robinson's Coefficient of Agreement

Robinson's coefficient of agreement A (Robinson, 1957, 1959) is expressed as

$$A = 1 - \frac{\sum_{i=1}^{n}(X_i - Z_i)^2 + \sum_{i=1}^{n}(Y_i - Z_i)^2}{\sum_{i=1}^{n}(X_i - \overline{Z})^2 + \sum_{i=1}^{n}(Y_i - \overline{Z})^2}, \quad (12)$$

where Z_i is the mean value of X_i and Y_i, $\overline{Z}$ is the mean value of $\overline{X}$ and $\overline{Y}$. Robinson's A is a bounded, non-dimensional and symmetric measure of agreement. The A value ranges potentially from 0 to 1, corresponding to the agreement from the worst to the best. But unlike Mielke's ρ that measures MSE between X and Y, Robinson's A measures the perpendicular distance between the point (X_i and Y_i) to the $Y = X$ line (Robinson, 1957, 1959).

Among the six groups of agreement measures, as summarized in Table 1, each one has advantages and disadvantages. For model validation, Willmott's method is more useful because it provides more detailed information regarding systematic and unsystematic errors. But Willmott's d is inadequate for measuring the agreement of data due to its asymmetry. Mielke's ρ and Robinson's A have advantages of boundedness, non-dimensionality and symmetry. But they are incapable of separating systematic and unsystematic errors.

TABLE 1. ADVANTAGES AND DISADVANTAGES OF THE MEASURES OF AGREEMENT

Measure of Agreement (X versus Y)	Advantage	Disadvantage
r and r^2	• Bounded • Dimensionless • Symmetric • r^2 states proportion of the variation explained by model	• Measuring linear covariation, not actual difference
MBE, MAE, and RMSE	• Measuring actual difference • Symmetric	• Not bounded • Dimensional
MAPE and RMSPE	• Dimensionless	• Asymmetric • Not bounded • Meaningless if X_i is negative • Unstable if X_i is small
Willmott's d	• Bounded • Dimensionless • Partitioning total error into systematic and unsystematic errors	• Asymmetric
Mielke's ρ	• Bounded • Dimensionless • Symmetric	• Unable to separate systematic and unsystematic errors
Robinson's A	• Bounded • Dimensionless • Symmetric	• Unable to separate systematic and unsystematic errors

A New Measure of Agreement: Agreement Coefficient

Criteria and Strategy

According to the previous discussion on the measures of agreement, we suggest that a good measure should meet four criteria: (a) Non-dimensional; (b) Bounded. Preferably, the measure ranges from 0 to 1. Thus, any values between 0 and 1 indicate the degree of agreement from complete disagreement to complete agreement; (c) Symmetric. Interchanging datasets X and Y should produce an identical result; and (d) Capable of separating systematic and unsystematic differences. Based on those criteria, we designed a measure of data agreement – agreement coefficient.

The strategy of the agreement coefficient between datasets X and Y is based on the evaluation of distances between the actual observation (X_i,Y_i), the 45° line and the regression line. The 45° line is line $Y = X$; the regression line is regression function $\hat{Y} = a + bX$ or $\hat{X} = c + dY$, estimated from datasets X and Y. Thus, there are three types of distances among the points and the two lines: (a) The distance between point (X_i,Y_i) and the 45° line is the difference between X_i and Y_i; (b) The distance between point (X_i,Y_i) and the linear regression line is the unsystematic difference between X_i and Y_i. The unsystematic difference is due to the random measurement error, which cannot be adjusted by the regression function; and (c) The distance between the 45° line and the linear

regression line is the systematic difference distance between X_i and Y_i. The systematic difference is attributed to the fixed difference between datasets X and Y. By applying the linear regression function between X and Y, dataset X can be transformed to dataset Y for the systematic difference portion, and conversely, Y can be transformed to X.

The difference between X and Y, however, has no upper limit and also is unit dependent. Thus, another quantity is needed to standardize the difference between X and Y, so as to make the agreement coefficient non-dimensional and bounded. This quantity is referred to as the potential difference, which is measured by the range of X and Y. The range of X and Y is the distance between mean values of X and Y, plus the distance between the mean values and the actual point (X_i,Y_i).

Product-difference is used to measure the distance between a point and a line, either a regression line or the 45° line. The geometric interpretation of the product-difference is illustrated in Figure 1. From point P (a,b) to line L, there is a horizontal distance QP, which is the distance from P (a,b) to Q (c,b), or $|a - c|$. There is also a vertical distance PR from P (a,b) to R (a,d), or $|b - d|$. We use $(|a - c|)(|b - d|)$, the product of the horizontal distance and the vertical distance, to measure the distance from P to L (Figure 1). For a special case that line L is the 45° line, the horizontal distance $|a - c|$ and the vertical distance $|b - d|$ are equal. Then, the product-difference between P and L becomes the squared difference of $|a - c|$, or $(a - c)^2$.

Agreement Coefficient

The agreement coefficient (AC) is defined as

$$AC = 1 - \frac{\sum_{i=1}^{n} (X_i - Y_i)^2}{\sum_{i=1}^{n} (|\overline{X} - \overline{Y}| + |X_i - \overline{X}|)(|\overline{X} - \overline{Y}| + |Y_i - \overline{Y}|)}, \quad (13)$$

where $\overline{X}$ and $\overline{Y}$ are the mean values of X and Y, respectively. The numerator of the main term of Equation 13 is sum of square difference (SSD) of X and Y, which indicates the degree of disagreement between X and Y. The denominator is the sum of potential difference (SPOD) used to standardize SSD. Thus, Equation 13 can by expressed in a simplified equation:

$$AC = 1 - \frac{SSD}{SPOD}. \quad (14)$$

The arithmetic mean of SSD is the mean square difference (MSD), given by:

$$MSD = \frac{1}{n} \sum_{i=1}^{n} (X_i - Y_i)^2. \quad (15)$$

Square root of MPD is the root mean square difference (RMSD), given by

$$RMSD = (MSD)^{1/2}. \quad (16)$$

SSD, MSD, and RMSD are knows as sum of SSE, MSE, and RMSE.

If the agreement between X and Y is perfect, i.e., $X_i = Y_i$ for $i = 1, 2, \ldots, n$, SSD = 0, and AC = 1. If $X_i = \overline{X}$ and $Y_i = \overline{Y}$ for all observations, SSD = SPOD, and AC = 0. Thus, AC is bounded below by 0 and above by 1. Because the units of SSD and SPOD cancels out in Equation 14, AC is dimensionless. Additionally, AC is symmetric, since the identical AC value is obtained if X and Y are interchanged.

The X-Y diagram in Figure 2a illustrates the graphic interpretation of SSD. On the diagram, the 45° line is $Y = X$ line. i.e., $X_i = Y_i$ $(i = 1, 2, \ldots n)$. For an observation (X_i,Y_i), there is a horizontal distance and a vertical distance from the point to the 45° line. The horizontal distance between point (X_i,Y_i) and the 45° line is the distance between (X_i,Y_i) and (Y_i,Y_i) on X axis, which equals $|X_i - Y_i|$. Similarly, the vertical distance between the point (X_i,Y_i) and the 45° line is the distance between (X_i,Y_i) and (X_i,X_i) on Y axis, which is also $|X_i - Y_i|$. The product of the horizontal distance $|X_i - Y_i|$ and the vertical distance $|X_i - Y_i|$ is $(X_i - Y_i)^2$, which measures the difference between X_i, and Y_i. The SSD, $\Sigma(X_i - Y_i)^2$, measures the squared difference between X_i and Y_i for all points $(i = 1, 2, \ldots n)$. Now we demonstrate an example of datasets X (6, 8, 9, 10, 11, 14) and Y (2, 3, 5, 5, 6, 8) and explain the calculation of the difference based on the geometric distance. For an example of the 3rd observation, $X_3 = 9$ and $Y_3 = 5$, the horizontal and vertical distances are $|X_3 - Y_3| = 4$, and the squared difference is $(X_3 - Y_3)^2 = 16$. For all the observations, SSD = 143, MSD = 23.833, and RMSD = 4.882.

The SPOD, the denominator of main term in Equation 13, represents the total range of X and Y, that includes the variability of X and Y, plus the difference between mean X and mean Y values. On the X-Y diagram shown in Figure 2b, the potential difference on X is the horizontal distance between (X_i,Y_i) and $(\overline{X},\overline{Y})$, plus the horizontal distance between $(\overline{X},\overline{Y})$ and the 45° line. The potential difference on Y is the vertical distance between (X_i,Y_i) and $(\overline{X},\overline{Y})$, plus the vertical distance between $(\overline{X},\overline{Y})$ and the 45° line. In other words, the potential difference on X is $(|\overline{X} - \overline{Y}| + |X_i - \overline{X}|)$ and on Y is $(|\overline{X} - \overline{Y}| + |Y_i - \overline{Y}|)$. Thus, the potential difference between X_i and Y_i is the product of $(|\overline{X} - \overline{Y}| + |Y_i - \overline{Y}|)$ and $(|\overline{X} - \overline{Y}| + |X_i - \overline{X}|)$. For all observations, SPOD equals $\Sigma(|\overline{X} - \overline{Y}| + |Y_i - \overline{Y}|)(|\overline{X} - \overline{Y}| + |X_i - \overline{X}|)$. In the above example, $\overline{X} = 9.667$ and $\overline{Y} = 4.833$. The potential difference for the 3rd observation is $(|X_i - \overline{X}| + |\overline{X} - \overline{Y}|)(|Y_i - \overline{Y})$

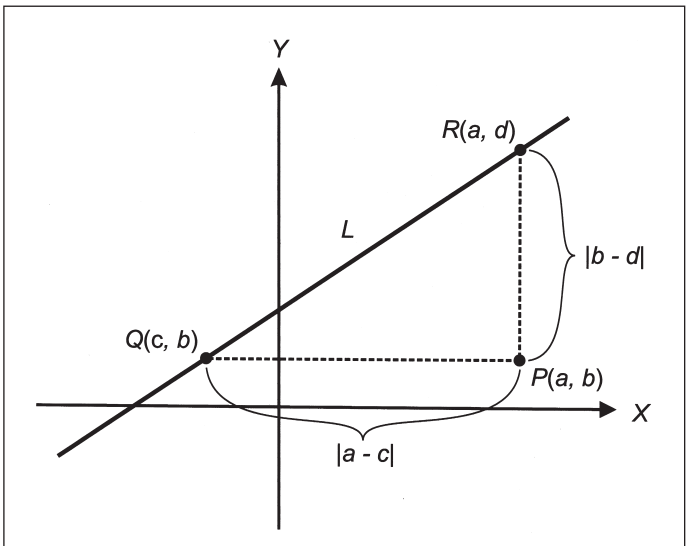

Figure 1. Illustration of the distance measures using product-difference. Distance from point $P(a,b)$ to line L is measured with the product of horizontal distance QP and vertical distance PR, i.e., $(|a - c|)(|b - d|)$.

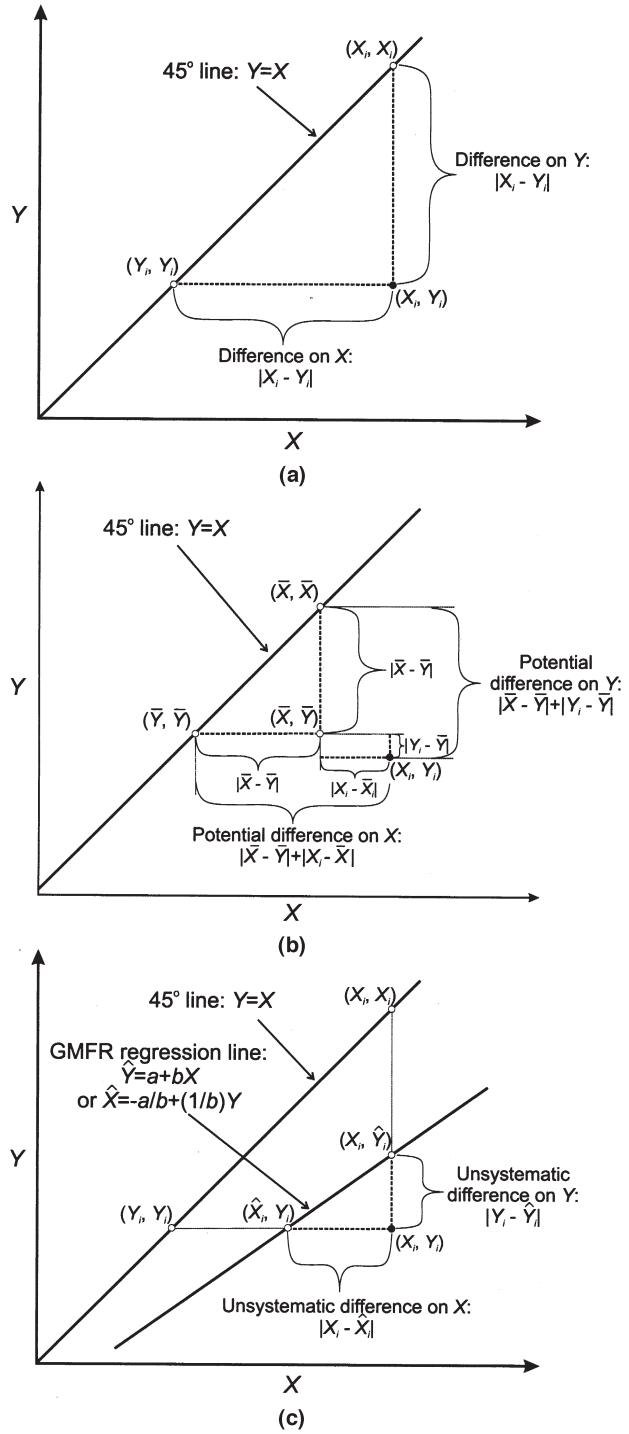

Figure 2. Measure of the difference between Datasets *X* and *Y*. The solid circle is an observation from datasets *X* and *Y*. The straight lines are the 45° line ($Y = X$) and the geometric mean functional relationship (GMFR) regression line. (a) The difference between *X* and *Y* is calculated based on the distance between the point (X_i,Y_i) and the 45° line on *X* and *Y* axes. (b) The potential difference is determined with the distance between (X_i,Y_i) and $(\overline{X},\overline{Y})$, and between $(\overline{X},\overline{Y})$ and the 45° line. (c) The unsystematic difference is measured by the distance between the point (X_i,Y_i) and the GMFR regression line on *X* and *Y* axes.

$+|\overline{X} - \overline{Y}|) = 5.5 \times 5 = 27.5$ For all observations, SPOD is 272.167, and finally, AC = 0.475.

Geometric Mean Functional Relationship Regression

To determine the systematic and unsystematic differences between *X* and *Y*, it is necessary to create a linear regression for *X* and *Y*. The distance between a given point (X_i,Y_i) and the regression line is the unsystematic difference of X_i and Y_i, and the distance between the 45° line and the linear regression line is the systematic difference. It should be emphasized that the least squares (LS) regression method used for regular regression analysis is not adequate for the data agreement analysis, because the LS method estimates the regression coefficients by minimizing the sum of squares of residuals, i.e., $\Sigma(Y_i - \hat{Y}_i)$. This method assumes that *Y* is subject to random error, but *X* is fixed (Draper and Smith 1998). If regression function $\hat{Y} = a + bX$ (or "*Y* versus *X* regression") is algebraically inverted, then $X = -a/b + (1/b)\hat{Y}$. Alternately, we can estimate the coefficients for regression $\hat{X} = c + dY$ (or "*X* versus *Y* regression") with the LS method. But the coefficients of $\hat{X} = c + dY$ with LS estimation are generally different from the coefficients obtained from the inverse regression function $X = -a/b + (1/b)\hat{Y}$ (i.e., $c \neq -a/b$, $d \neq 1/b$). Using the previous example, the LS estimation generates a *Y* versus *X* regression as $\hat{Y} = -2.589 + 0.768X$ (or in inverse format, $X = 3.371 + 1.302\hat{Y}$), and an *X* versus *Y* regression as $\hat{X} = 3.59 + 1.256Y$. The two regression functions, *Y* versus *X* and *X* versus *Y*, are not coincident. Therefore, LS regression is asymmetric since only errors on *Y* are taken into consideration in the model estimation and *X* is assumed error-free.

Willmott's method to estimate systematic and unsystematic errors is based on LS regression $\hat{Y} = a + bX$. When $\hat{Y}$ is obtained, the MSE_s and MSE_u can be calculated by following Equation 8 and Equation 9. LS regression $\hat{Y} = a + bX$ causes an asymmetric characteristic for Willmot's MSE_s and MSE_u. That is the reason why MSE_s and MSE_u produce different values when *X* and *Y* are interchanged. The use of Willmott's indices for model validation is deemed reasonable if actual observed and model-predicted values are compared. Willmott (1981) assumed that observed values are error-free when they are compared to model-predicted values. If the comparison is between two actually measured datasets, both datasets may be equally subject to measurement errors. We need to find a different regression method that deals with errors in both *X* and *Y*, thus $\hat{Y} = a + bX$ is symmetric, or invertible. The geometric mean functional relationship (GMFR) model (Ricker, 1980; Laws and Archie, 1981; Draper and Smith, 1998) is a symmetric regression model, with an assumption that both *X* and *Y* are subject to errors. The GMFR model is expressed in the same way as LS regression model

$$\hat{Y} = a + bX. \tag{17}$$

where *Y* and *X* are variables both subject to errors. But the coefficients are estimated with the equations that are different from LS estimation:

$$b = \pm \left(\frac{\sum_{i=1}^{n} (Y_i - \overline{Y})^2}{\sum_{i=1}^{n} (X_i - \overline{X})^2} \right)^{\frac{1}{2}}; \tag{18}$$

$$a = \overline{Y} - b\overline{X}. \quad (19)$$

Here, the sign of b is same as the correlation coefficient of X and Y. From Equation 17, we get the inverse function $X = -a/b + (1/b)\hat{Y}$. The coefficients are exactly the same as those of $\hat{X} = c + dY$ that we estimate for X versus Y regression using the GMFR method (that is, $-a/b = c$, $1/b = d$). On the Y versus X plot (Figure 2c), the fitted GMFR regression lines for $\hat{Y} = a + bX$ and $\hat{X} = c + dY$ are coincident. In the earlier example, the GMFR regressions for X and Y are $\hat{Y} = -2.727 + 0.782X$ and $\hat{X} = 3.486 + 1.279Y$. The two regression functions are identical, therefore algebraically invertible. GMFR regression is suggested in agreement analysis, since both datasets X and Y are considered being subject to measurement errors. Even if both X and Y contain negligible measurement error, GMFR regression is preferable to LS regression because X and Y are treated equally.

Systematic and Unsystematic Differences

Since $\hat{X}$ and $\hat{Y}$ are calculated from the GMFR regression, it is convenient to determine the systematic and unsystematic differences between X and Y. The unsystematic differences can be obtained by computing the sum of product-difference (SPD) and mean product-difference (MPD). The unsystematic sum of product-difference (SPD_u) is defined as

$$SPD_u = \sum_{i=1}^{n} (|X_i - \hat{X}_i|)(|Y_i - \hat{Y}_i|), \quad (20)$$

and the unsystematic mean product-difference (MPD_u) is

$$MPD_u = \frac{1}{n}(SPD_u). \quad (21)$$

Because the SSD can be partitioned into the SPD_u and the systematic sum of product-difference (SPD_s), the SPD_s can be obtained by

$$SPD_s = SSD - SPD_u. \quad (22)$$

The mean product-difference (MPD_s) is derived by the arithmetic mean of SPD_s, as

$$MPD_s = \frac{1}{n}(SPD_s). \quad (23)$$

The systematic square root of mean product-difference ($RMPD_s$) and unsystematic square root of mean product-difference ($RMPD_u$) are further given by

$$RMPD_s = (MPD_s)^{1/2}, \quad (24)$$

and

$$RMPD_u = (MPD_u)^{1/2}. \quad (25)$$

On the X-Y diagram in Figure 2c, the regression line $\hat{Y} = a + bX$, estimated with the GMFR method, coincides with regression line $\hat{X} = c + dY$. For an observation (X_i, Y_i), there is a horizontal distance from it to regression line $\hat{X} = c + dY$. The distance between (X_i, Y_i) and $(\hat{X}_i, Y_i)$, i.e., $|X_i - \hat{X}_i|$, is the unsystematic difference on X. Similarly, there is a vertical distance from (X_i, Y_i) to line $\hat{Y} = a + bX$. The distance between (X_i, Y_i) and $(X_i, \hat{Y}_i)$, or $|Y_i - \hat{Y}_i|$, is the unsystematic difference on Y. The product of $|X_i - \hat{X}_i|$ and $|Y_i - \hat{Y}_i|$ is used to indicate the unsystematic difference on both X and Y. For the 3rd observation in the example demonstrated above, $X_3 = 9$, $Y_3 = 5$, $\hat{Y}_3 = 9.880$, $\hat{X}_3 = 4.312$. The unsystematic difference on X is $|X_i - \hat{X}_i| = 0.880$; the unsystematic difference on Y is $|Y_i - \hat{Y}_i| = 0.688$; the product of the two differences is 0.605. For all the observations, $SPD_s = 141.942$, $MPD_s = 23.657$, $RMPD_s = 4.864$, $SPD_u = 1.062$, $MPD_u = 0.177$, and $RMPD_u = 0.420$.

Systematic and Unsystematic Agreement Coefficients

The agreement coefficient evaluates the total agreement, which includes the systematic and unsystematic agreement between two datasets. We can use systematic agreement coefficient (AC_s) and unsystematic agreement coefficient (AC_u) to measure the systematic and unsystematic agreement, respectively. AC_s and AC_u are calculated by replacing SSD in Equation 14 by SPD_s and SPD_u, given by

$$AC_s = 1 - \frac{SPD_s}{SPOD}, \quad (26)$$

and

$$AC_u = 1 - \frac{SPD_u}{SPOD}. \quad (27)$$

AC_s indicates the standardized data agreement between X and Y for the systematic portion, whereas AC_u measures the unsystematic portion. In the above example, AC_u and AC_s are 0.996 and 0.478, respectively.

In summary, AC, AC_s, AC_u, and the associated parameters (RMSD, $RMPD_s$, and $RMPD_u$) are suggested to indicate data agreement or disagreement. In addition, MPD_s/MSD and MPD_u/MSD expressed in percentage are useful because they measure the proportion of the systematic and unsystematic differences over the total difference. As a supplemental material, a SAS program including the SAS output for calculating the agreement coefficient is provided on the American Society for Photogrammetry and Remote Sensing's website (**www.asprs.org**).

Limitations for the Agreement Coefficient

The agreement coefficient was designed to quantify the degree of agreement or deviation between two matched datasets. The coefficient, however, does have some constraints when it is used for data comparison. First, the agreement coefficient requires that both of compared datasets are subject to about equal measurement errors. If the quality of one dataset is verified superior over the other, use of Willmott's index of agreement is more appropriate. Second, the agreement coefficient is not capable of detecting data quality. To determine the data quality, a separate and independent dataset is needed as a reference. The reference data should be collected with a better instrument or technique, for example, finer spectral and radiometric resolutions, better calibration and geometric correction, or even with ground truth measurements. Willmott's index of agreement can be used to discover the more accurate dataset based on the ancillary reference data. The third limitation is associated with the measure of systematic and unsystematic differences. The relationship between two datasets for comparison is constrained to be linear, because the GMFR regression requires a linear relationship between X and Y (Draper and Smith, 1998). Further studies are needed on nonlinear relationships, such as polynomial (order ≥ 2) and exponential relation patterns. Finally, the agreement coefficient is designed suitably only for the image comparison for quantitative (either discrete or continuous) variables. If categorical variables are involved in the image comparison, measures of agreement for categorical data should be applied (Congalton and Green, 1999; Pontius, 2002; Pontius and Suedmeyer, 2004).

The Use of Agreement Coefficient

$$Y = a + bX + e, \tag{28}$$

Agreement Coefficient Applied to Hypothetical Data

To demonstrate the performance of the agreement coefficient, we simulated and compared four pairs of hypothetical datasets. The hypothetical dataset X comprises a set of random numbers generated with normally distributed values, with mean of 350 and standard deviation of 100. Dataset Y was generated from a linear function of X

where a and b are the constants, e is the normally distributed random value. Four sets of Y were obtained by changing the values of a, b, and e in the transformation. The hypothetical data were converted to 20×10 pixels images (Figure 3), and each pair of the images represents a special case of difference between X and Y, as describe bellow:

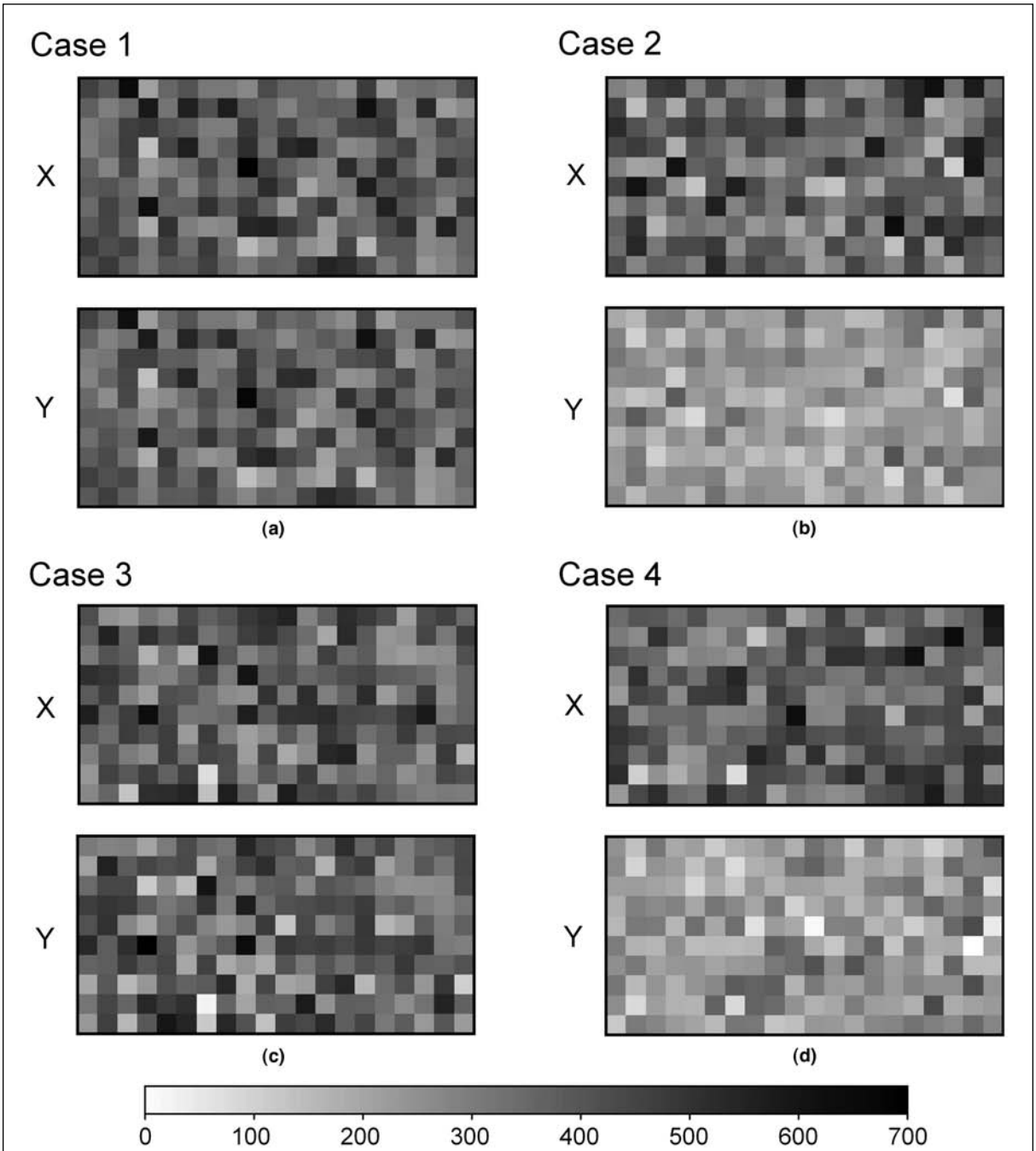

Figure 3. Four pairs of images for data agreement analysis. Each pair shows a special case regarding the systematic and unsystematic differences between datasets X and Y. (a) Case 1: Low systematic difference and low unsystematic difference. (b) Case 2: High systematic difference and low unsystematic difference. (c) Case 3: Low systematic difference and High unsystematic difference. (d) Case 4: High systematic difference and high unsystematic difference.

- Case 1: Low systematic difference and low unsystematic difference. The two images are very similar. Each pair of pixels from the two images have almost equal values, with very small random difference.
- Case 2: High systematic difference and low unsystematic difference. The two images show great differences between the corresponding pixels. Although all pixels in X are higher than Y, their difference is systematic (higher X value corresponds to higher Y value).
- Case 3: Low systematic difference and high unsystematic difference. The pixel values between the two images are very similar, but there are great random differences.
- Case 4: High systematic difference and high unsystematic difference. The pixel values of X are consistently higher than Y, and the random differences between X and Y are also high.

Plotting Y versus X is an intuitive way to demonstrate and understand the systematic and unsystematic differences between X and Y (Figure 4). The agreement analysis further provides a quantitative summary of the data agreement. The summary statistics and the parameters of agreement analysis for the hypothetical datasets are presented in Table 2. In Case 1, X and Y match well and r^2 is high (0.988). RMSD, $RMPD_s$, and $RMPD_u$ values are low (17.0, 13.8, and 10.0, respectively). AC, AC_s, and AC_u are high (0.972, 0.982, and 0.990, respectively). Case 2 shows very a high r^2 (0.976). But $RMPD_s$ (133.6) is much higher than $RMPD_u$ (12.3). Although AC_s is low (0.543), AC_u is high (0.996). Thus, the systematic difference contributes a greater amount of the data deference. For this case, there is a noticeable difference between r^2 and AC (0.539). In fact, AC value indicates the departure of the points from the 45° line, but r^2 measures the closeness of the points about the regression line. In Case 3, $RMPD_s$ is much lower (16.1) than $RMPD_u$ (51.7). AC_s and AC_u are 0.978 and 0.774, respectively. AC (0.752) is close to r^2 (0.765), due to the closeness of the regression line and the 45° line. Case 4 has high $RMPD_s$ and $RMPD_u$ values (156.4 and 137.8), and low AC, AC_s, and AC_u values (0.442, 0.567, and 0.872, respectively).

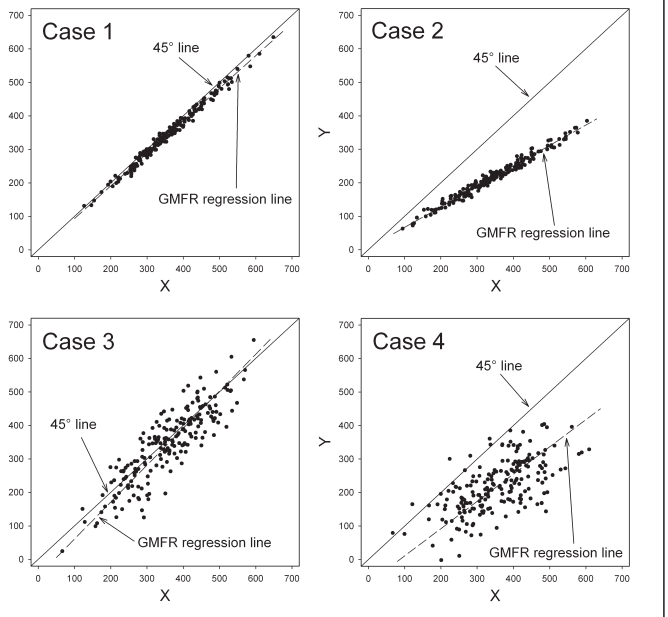

Figure 4. Scatterplots of simulated datasets Y and X for four spatial cases.

Agreement Coefficient Applied to Satellite Sensor Data

For the second set of examples, we compared the 2002 NDVI data derived from NOAA-16 AVHRR and Terra MODIS sensors. The two datasets are the 1 km resolution 16-day composite images for the conterminous United States. Although the NDVI products from the two sensors were corrected for molecular scattering (including water vapor), and ozone absorption (Defelice, 2003), there are some differences regarding the sensor systems and data processing techniques: (a) the AVHRR red (585 to 680 nm) and near-infrared (730 to 980 nm) bands are much wider than the MODIS red (620 to 670 nm) and near-infrared (841 to 876 nm) bands; (b) AVHRR images are resampled from the data with resolution of 1.1 km at nadir (the resolution becomes coarser with increasing viewing angle off-nadir). MODIS images are integrated from the original data at 250 m resolution; (c) Equator crossing time for the NOAA-16 platform is 1400 local standard time and for the Terra platform is 1035 local standard time; and (d) The AVHRR uses the maximum value composite technique (Eidenshink, 1992), and the MODIS uses the constrained-view angle – maximum value composite (Huete *et al.*, 2002).

Three 16-day composite intervals (Figure 5) were selected for the comparison from the 2002 AVHRR and MODIS NDVI data: intervals 8 (22 April through 07 May), 13 (11 through 26

TABLE 2. AGREEMENT ANALYSIS FOR HYPOTHETICAL DATASETS

Case	Case 1	Case 2	Case 3	Case 4
n	200	200	200	200
Mean	X: 361.0	X: 342.9	X: 357.0	X: 354.8
	Y: 347.4	Y: 215.2	Y: 344.4	Y: 218.2
Difference ($\overline{X} - \overline{Y}$)	13.6	127.7	12.6	136.6
Standard deviation	X: 92.5	X: 101.6	X: 98.6	X: 96.1
	Y: 89.9	Y: 62.2	Y: 108.8	Y: 78.2
r^2	0.988	0.976	0.765	0.391
AC	0.972	0.539	0.752	0.442
AC_s	0.982	0.543	0.978	0.567
AC_u	0.990	0.996	0.774	0.872
RMSD	17.0	134.2	54.1	156.4
$RMPD_s$	13.8	133.6	16.1	137.8
$RMPD_u$	10.0	12.3	51.7	74.8
MPD_s/MSD (%)	65.8	99.2	8.5	77.1
MPD_u/MPD (%)	34.2	0.8	91.5	22.9
GMFR regression (Y versus X)	$\hat{Y} = -3.466 + 0.972X$	$\hat{Y} = 5.554 + 0.611X$	$\hat{Y} = -49.309 + 1.103X$	$\hat{Y} = -70.401 + 0.813X$
GMFR regression (X versus Y)	$\hat{X} = 3.566 + 1.029Y$	$\hat{X} = -9.083 + 1.635Y$	$\hat{X} = 44.709 + 0.907Y$	$\hat{X} = 86.554 + 1.229Y$

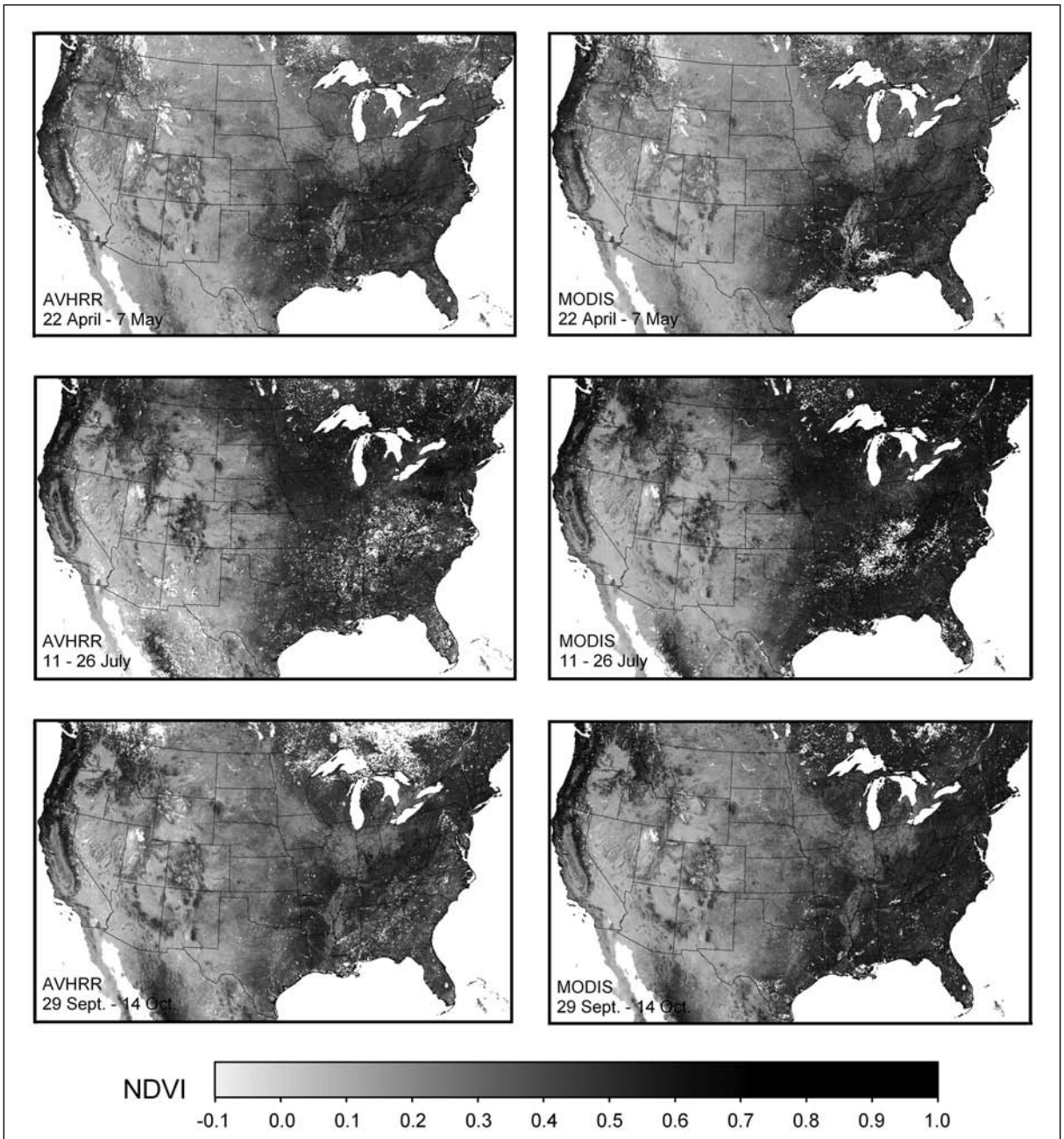

Figure 5. AVHRR-NDVI and MODIS-NDVI Images for the conterminous United States. The data are the 16-day Composites from 2002. Water and cloud pixels (in white) are excluded from the images.

July), and 18 (29 September through 14 October). Water areas that derived from the National Land Cover Data Set (Vogelmann *et al.*, 2001), and cloud areas extracted from the Cloud From AVHRR data and the MODIS Quality Assurance data were masked in the NDVI images. To reduce the image random noises caused by pixel misregistration, we applied 5×5 pixels median-filter to the images. A pixel-to-pixel comparison was made for each pair of the images. Thus, all pixels within the United States boundary, with exclusion of water and cloud pixels, were used for the agreement analysis.

Figure 6 displays the density scatterplots of MODIS versus AVHRR for the three pairs of NDVI images. Linear regression functions are found best to describe the relationship between AVHRR-NDVI and MODIS-NDVI for all the three intervals. The GMFR regression line deviates slightly from the 45° line, especially for higher NDVI values. The plots indicate that both systematic and unsystematic differences exist, and the unsystematic differences are higher. The results of agreement analysis are presented in Table 3. For all the three intervals, MODIS has slightly higher NDVI values than AVHRR. The r^2 values are high (0.91 to 0.94) for the three intervals, and AC values are also high (0.90 to 0.94). Again, note that the r^2 value does not reflect the true data bias between AVHRR-NDVI and MODIS-NDVI, because they only accounts for the data deviation about the regression line. However, if the systematic difference is low, or the regression line is close to the 45° line, the r^2 and the AC values are close. The systematic and unsystematic differences, indicted by $RMPD_s$ and $RMPD_u$, are 0.02 to 0.05 and 0.05 to 0.06. The unsystematic difference accounts for a higher

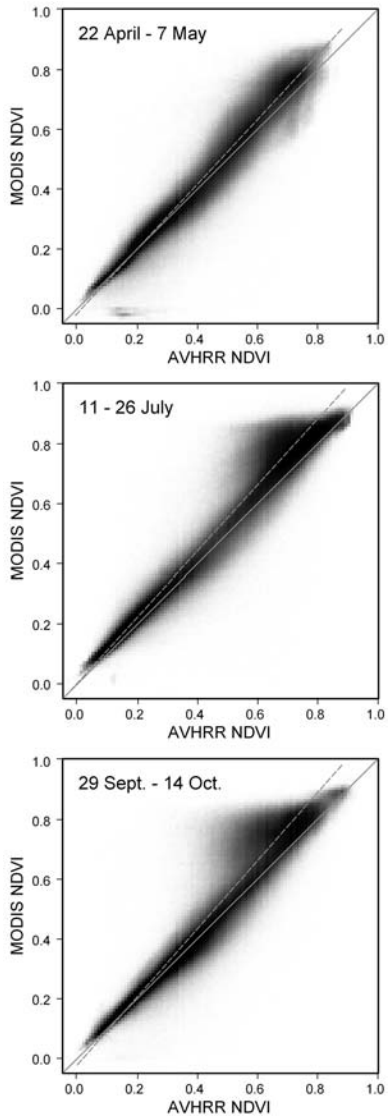

Figure 6. Density scatterplots of AVHRR NDVI versus MODIS NDVI. The gray solid line is the 45° line and the gray dashed line is the GMFR regression line. Numbers of samples (pixels) are 7,521,372, 7,309,410, and 7,553,976 for the three 16-day intervals. The darker area on the plot indicates higher density of the sample points.

proportion (61 to 81 percent) of total difference, while the systematic deference is 19 to 39 percent. Therefore, for all the three composite intervals, AC_s values (0.988, 0.972, and 0.970) are higher than AC_u (0.949, 0.956, and 0.929).

The systematic difference between the AVHRR and MODIS NDVI can be eliminated by applying GMFR regression to the data. That is, AVHRR-NDVI can be transformed to MODIS-NDVI, and vice versa, based on the regression function. The linear transform using the regression function, however, is unable to reduce the unsystematic difference between AVHRR and MODIS data. In this paper, we just demonstrate some simplified examples from AVHRR and MODIS data on how to use the agreement coefficient to compare images. A complete comparison for AVHRR and MODIS NDVI data using the agreement analysis method will be presented in a separate paper.

Conclusions

The agreement coefficient was developed to meet the needs for evaluation of image agreement. The method of agreement analysis were designed based on the data difference among actual observations, $Y = X$ line and regression line. Unlike the traditional agreement measures (e.g., r, r^2, MAE and RMSE) and the model-validation measures (e.g., Willmott's and Mielke's measures), this new agreement coefficient is non-dimensional, bounded, symmetric, and capable of separating systematic and unsystematic differences. Non-dimensionality and boundedness are critical for cross-data comparison. Symmetry property assures an identical result when X and Y are interchanged. The systematic difference can be adjusted by applying a linear function; but the unsystematic difference is the random error, which is not able to be eliminated by the linear transformation between X and Y.

In this study, the agreement coefficient was applied to the hypothetical data and the real remotely sensed data and was shown effective in evaluating image agreement, including systematic and unsystematic agreement. This agreement analysis method is especially useful for comparing remote sensed datasets derived from different sensors. The new technique, however, is not only limited to the use in remote sensing, it can be furthermore used to compare the quantitative data in other geographical, environmental and climatologic studies.

Acknowledgments

This study was supported in part by the National Polar-orbiting Operational Environmental Satellite System Integrated Program Office's Internal Government Studies Program. We would like to thank Stephen Stehman, Jerry Sullivan, and the three anonymous reviewers for their valuable comments on the paper manuscript.

References

Bogaert, J., L. Zhou, C.J. Tucker, R.B. Myneni, and R. Ceulemans, 2002. Evidence for a persistent and extensive greening trend in Eurasia inferred from satellite vegetation index data, *Journal of Geophysical Research*, 107(D11), doi:10.1029/2001JD001075.

Congalton, R., and K. Green, 1999. *Assessing the Accuracy of Classification of Remotely Sensed Data: Principle and Practices*, Lewis Publishers, Boca Raton, Florida, 137 p.

Defelice, T.P., D. Lloyd, D.J. Meyer, T.T. Baltzer, and P. Piraino, 2003. Water vapour correction of the daily 1 km AVHRR global land dataset: Part I validation and use of the water vapour input field, *International Journal of Remote Sensing*, 24(11):2365–2375.

Draper, N.R., and H. Smith, 1998. *Applied Regression Analysis*, John Wiley & Sons, Inc., New York, 706 p.

Eidenshink, J.E., 1992. The 1990 Conterminous U.S. AVHRR data set, *Photogrammetric Engineering & Remote Sensing*, 58(6):809–813.

Gallo, K., L. Ji, B. Reed, J. Dwyer, and J. Eidenshink, 2004. Comparison of MODIS and AVHRR 16-day normalized difference vegetation index composite data, *Geophysical Research Letters*, 31:L07502, doi:10.1029/2003GL019385.

Gao, B., 2000. A practical method for simulating AVHRR-consistent NDVI data series using narrow MODIS channels in the 0.5–1.0 μm spectral range, *IEEE Transactions on Geoscience and Remote Sensing*, 38(4):1969–1975.

Gitelson, A.A., and Y.J. Kaufman, 1998. MODIS NDVI optimization to fit the AVHRR data series – Spectral considerations, *Remote Sensing of Environment*, 66:343–350.

Goetz, S.J., 1997. Multi-sensor analysis of NDVI, surface temperature and biophysical variables at a mixed grassland site, *International Journal of Remote Sensing*, 18(1):71–94.

Goward, S.N., P.E. Davis, D. Fleming, L. Miller, and J.R. Townshend, 2003. Empirical comparison of Landsat 7 and IKONOS multi-

TABLE 3. AGREEMENT ANALYSIS FOR AVHRR AND MODIS NDVI

16-day Interval	22 April to 7 May	11 to 26 July	29 September to 14 October
n (pixel)	7521372	7309410	7553976
Mean	A: 0.420	A: 0.505	A: 0.460
	M: 0.435	M: 0.547	M: 0.491
Difference ($\overline{A} - \overline{M}$)	−0.015	−0.042	−0.031
Standard deviation	A: 0.195	A: 0.235	A: 0.199
	M: 213	M: 0.255	M: 0.226
r^2	0.944	0.943	0.913
AC	0.937	0.930	0.898
AC_s	0.988	0.972	0.970
AC_u	0.949	0.956	0.929
RMSD	0.054	0.075	0.075
$RMPD_s$	0.024	0.047	0.041
$RMPD_u$	0.049	0.059	0.063
MPD_s/MSD (%)	19.0	38.8	29.4
MPD_u/MSD (%)	81.0	61.2	70.6
GMFR regression (M versus A)	$\hat{M} = -0.024 + 1.093A$	$\hat{M} = -0.0004 + 1.085A$	$\hat{M} = -0.031 + 1.135A$
GMFR regression (A versus M)	$\hat{A} = 0.022 + 0.914M$	$\hat{A} = 0.0004 + 0.922M$	$\hat{A} = 0.027 + 0.881M$

Note: A = AVHRR-NDVI; M = MODIS-NDVI.

spectral measurements for selected Earth Observation System (EOS) validation sites, *Remote Sensing of Environment*, 88:80–90.

Huete, A., K. Didan, T. Miura, E.P. Rodriguez, X. Gao, and L.G. Ferreira, 2002. Overview of the radiometric and biophysical performance of the MODIS vegetation indices, *Remote Sensing of Environment*, 83:195–213.

Laws, E.A., and J.W. Archie, 1981. Appropriate use of regression analysis in marine biology, *Marine Biology*, 65:13–16.

Loveland, T.R., T.L. Sohl, S.V. Stehman, A.L. Gallant, K.L. Sayler, and D.E. Napton, 2002. Strategy for estimating the rates of recent United States land-cover changes, *Photogrammetric Engineering & Remote Sensing*, 68(10):1091–1099.

Mielke, P.W., Jr., 1984. Meteorological applications of permutation techniques based on distance functions, *Handbook of Statistics, Volume 4* (P.R. Krishnaiah and P.K. Sen, editors), Elsevier, New York, pp. 813–830.

Mielke, P.W., Jr., 1991. The application of multivariate permutation methods based on distance functions in the earth science, *Earth-Science Review*, 31:55–71.

Mielke, P.W., Jr., K.J. Berry, C.W. Landsea, and W.M. Gray, 1996. Artificial skill and validation in meteorological forecasting, *Weather and Forecasting*, 11:153–169.

Mielke, P.W., Jr., K.J. Berry, C.W. Landsea, and W.M. Gray, 1997. A single-sample estimate of shrinkage in meteorological forecasting, *Weather and Forecasting*, 12:847–858.

Pontius, R., 2002. Statistical methods to partition effects of quantity and location during comparison of categorical maps at multiple resolutions, *Photogrammetric Engineering & Remote Sensing*, 68(10):1041–1049.

Pontius, R., and B. Suedmeyer, 2004. Components of agreement in categorical maps at multiple resolutions, *Remote Sensing and GIS Accuracy Assessment* (R.S. Lunetta and J.G. Lyon, editors), CRC Press, Boca Raton, Florida, pp. 233–251.

Ricker, W.E., 1980. Computation and use of central lines, *Canadian Journal of Zoology*, 62:1897–1905.

Ramanathan, R., 1995. *Introductory Econometrics with Applications* (3rd edition), Harcourt Brace College Publishers, Fort Worth, 800 p.

Robinson, W.S., 1957. The Statistical Measurement of agreement, *American Sociological Review*, 22(1):17–25.

Robinson, W.S., 1959. The Geometric interpretation of agreement, *American Sociological Review*, 24(3):338–345.

Steven, M.D., T.J. Malthus, F. Baret, H. Xu, and M.J. Chopping, 2003. Intercalibration of vegetation indices from different sensor systems, *Remote Sensing of Environment*, 88:412–422.

Teillet, P.M., K. Staenz, and D.J. William, 1997. Effects of spectral, spatial, and radiometric characteristics on remote sensing vegetation indices of forested regions, *Remote Sensing of Environment*, 61:139–149.

Teillet, P.M., J.L. Barker, B.L. Markham, R.R. Irish, G. Fedosejevs, and J.C. Storey, 2001. Radiometric cross-calibration of the Landsat-7 ETM+ and Landsat-5 TM sensors based on tandem data sets, *Remote Sensing of Environment*, 78:39–54.

Vogelmann, J.E., S.M. Howard, L. Yang, C.R. Larson, B.K. Wylie, and N. Van Driel, 2001. Completion of the 1990s National Land Cover Data Set for the conterminous United States from Landsat Thematic Mapper Data and ancillary data sources, *Photogrammetric Engineering & Remote Sensing,* 67(6):650–662.

Willmott, C.J., 1981. On the validation of models, *Physical Geography,* 2(2):184–194.

Willmott, C.J., 1982. Some comments on the evaluation of model performance, *Bulletin American Meteorology Society,* 63(11): 1309–1313.

Willmott, C.J., S.G. Ackleson, R.E. Davis, J.J. Feddema, K.M. Klink, D.R. Legates, J. O'Donnell, and C.M. Rowe, 1985. Statistics for the evaluation and comparison of models, *Journal of Geophysical Research,* 90(C5):8995–9005.

Zhou, L., C.J. Tucker, R.K. Kaufmann, D. Slayback, N.V. Shabanov, and R.B. Myneni, 2001. Variations in northern vegetation activity inferred from satellite data of vegetation index during 1981 to 1999, *Journal of Geophysical Research,* 106(D17): 20069–20083.

Zhou, L., R.K. Kaufmann, Y. Tian, R.B. Myneni, and C.J. Tucker, 2003. Relation between interannual variations in satellite measures of northern forest greenness and climate between 1982 and 1999, *Journal of Geophysical Research,* 108(D1):4004, doi:10.1029/2002JD002510.

(Received 03 February 2005; accepted 16 May 2005; revised 31 May 2005)