# Statistical Methods

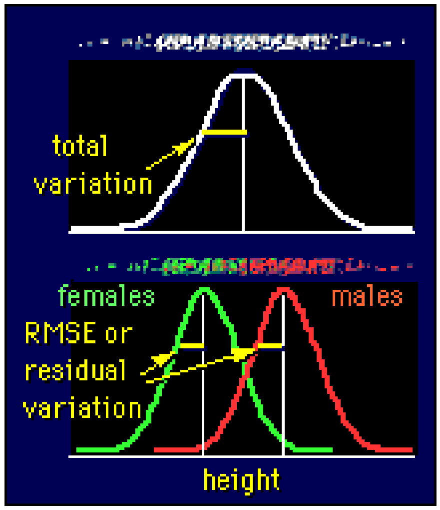
CompassAA produces several summary statistics to use as a measure of accuracy. Each of the statistics produced provides a different metric by which to judge and assess the data. This section will provide information on each of the statistical metrics provided by CompassAA. This section will also describe what each metric means and how it applies to a given project.

*Root Mean Square Error (RMSE)*

The RMSE is used to describe accuracy encompassing both random and systematic errors. RMSE is the square root of the square of the difference between a true test point and an interpolated test point divided by the total number of test points in the arithmetic mean. CompassAA utilizes the following equation (1) for calculating the RMSE:

$$RMSX = \sqrt{((x_1^2 + x_2^2 + ... + x_n^2)/n)} \qquad (1)$$

Example 1: Suppose you have heights for a group of females and males. If you analyze the data without regard to the sex of the subjects, the measure of spread you get will be the total variation. Most statistical programs can take into account the sex of each subject, work out the arithmetic means for the boys and the girls, and then derive a single standard deviation that will do for both boys and girls. That single standard deviation is the RMSE. In CompassAA, RMSE is used with the ΔX and ΔY values.




**Example 1.** Total variation compared to RMSE.

*Minimum ΔX and Minimum ΔY*

ΔX and ΔY represent the change in value between the X(Y)-coordinate location in the photo and X(Y)-coordinate location in the survey. The minimum of each simply represents the smallest value out of all of the analyzed locations. This component of the statistics can be used to show which user entered locations most closely correspond to the surveyed points.

*Maximum ΔX and Maximum ΔY*

As with the minimum ΔX and minimum ΔY, the ΔX and ΔY still represent the change in value between the user-entered coordinate location and the surveyed coordinate location. The maximum value of each represents the largest value out of all of the analyzed locations. Maximum values can be used to show which user entered locations most closely correspond to the surveyed points.

*Mean ΔX and Mean ΔY*

The statistics of the mean ΔX and mean ΔY are intended to show the average values of the ΔX and ΔY. Again, ΔX and ΔY show the change in value between the user-entered coordinate locations and the survey coordinate locations. The mean values can be used to show the average off-sets and of all of the analyzed locations providing valuable information about the distribution of the data.

*CE90 and CE 95*

CE stands for circular error probable. It is defined as a circle, centered about the mean, whose boundary is expected to include a certain percentage of the population within it. In the case of CE90, 90% of the population is expected to be included within the circular radius and in the case of CE95, 95% of the population is expected to be included within the circular radius.

In CompassAA, the horizontal positional error of an object can be represented by a random variable pair, *(x, y)*. The random variables *x* and *y* correspond to the error encountered in the X (longitude) and Y (latitude) directions respectively. The error can be considered as the deviation of the measured values from the true values. The two random variables can be assumed to be independent, with a Gaussian distribution and zero mean. The joint probability density distribution for these random variables (*x*, *y*) is given by equation (2). Rearranging equation (2) results in (3).

$$p(x,y) = \frac{1}{2\pi\sigma_x\sigma_y} e^{-\frac{1}{2}\left(\frac{x^2}{\sigma_x^2}+\frac{y^2}{\sigma_y^2}\right)} \qquad (2)$$

$$-2\ln[p(x,y)2\pi\sigma_x\sigma_y] = \left(\frac{x^2}{\sigma_x^2}+\frac{y^2}{\sigma_y^2}\right) \qquad (3)$$

As observed in equation (3) for a given value (*x, y*) the probability density function represents the square of the radius of the circle assuming that variances ( $\sigma_x$ and *y* $\sigma_y$ ) in both the dimensions are equal. The probability for an error random variable pair (*x, y*) to be contained within a circle of radius R can be defined by the circular error probability function *P(R).* The circular error probability function can be derived from equation (3) and is detailed in [2]. A condensed form for *P(R)* for the case when $\sigma_x$ and *y* $\sigma_y$ are equal is given by equation (4)

$$P(R) = 1 - e^{-\frac{R^2}{2\sigma_c^2}} \qquad (4)$$

where R is the radial distance.

For CE90, the National Map Accuracy Standard (NMAS) specifies that the 90% of well-defined points in an image or map should fall within a certain radial distance R. For CE95, the National Map Accuracy Standard (NMAS) specifies that the 95% of well defined points in an image or map should fall within a certain radial distance R. Therefore, substituting the left hand side of (4) with 0.90 will yield the horizontal accuracy standard as specified by NMAS which is given by equation (5).

$$CE_{90} = 2.1460\sigma_c \qquad (5)$$

where $\sigma_x = \sigma_y = \sigma_c$

Similarly, for CE95 substituting 0.95 with 0.9 then simplifying yields.

$$CE_{95} = 2.4477\sigma_c$$

The calculation for $\sigma_x$ is shown below.

$$\sigma_x = \sqrt{\frac{\Sigma(x_{image} - x_{realworld})^2}{n}} \qquad (6)$$

where $X_{image}$ and $X_{realworld}$ are the coordinates of the control points measured from the image and real world respectively, and *n* is the number of such control points. $\sigma_y$ is calculated similar to (6).

For cases where $\sigma_x$ and $\sigma_y$ are not equal, the error distribution takes on a more elliptical shape rather than being truly circular. However, it is shown in [2] that a Gaussian circular distribution can be still substituted for the elliptical distribution for certain $\frac{\sigma\, min}{\sigma\, max}$ ratios, where $\sigma_{min}$ is the minimum value between $\sigma_x$ and $\sigma_y$ , and $\sigma_{max}$ is the maximum value between $\sigma_x$ and $\sigma_y$ .

For cases where $\sigma_x$ and $\sigma_y$ are not equal and $\frac{\sigma\, min}{\sigma\, max}$ ratio is between 0.6 and 1.0, [1] shows that $\sigma_c$ is estimated by a linear combination of $\sigma_x$ and $\sigma_y$ as given by equation (7).

$$\sigma_c = 0.5222\sigma_{\min} + 0.4778\sigma_{\max} \qquad (7)$$

CompassAA™

In equation (7), $\sigma_{min}$ is the minimum value between $\sigma_x$ and $\sigma_y$, and $\sigma_{max}$ is the maximum value between $\sigma_x$ and $\sigma_y$. A further approximation of (7) is given in equation (8), which was adopted by NSSDA (Federal Geographic Data Committee 1988), the United States' standard for spatial data.

$$\sigma_c = 0.5(\sigma_{\min} + \sigma_{\max}) \qquad (8)$$

For cases where $\sigma_x$ and $\sigma_y$ are not equal and $\frac{\sigma\, min}{\sigma\, max}$ ratio is between 0.2 and 0.6, $\sigma_c$ is estimated using an interpolated value from statistical data that relates $\frac{\sigma\, min}{\sigma\, max}$ to $\frac{\sigma\, c}{\sigma\, max}$

*Skew*

Skew is a measure of symmetry, or more precisely, lack thereof. The skew for a normal distribution is zero, and any symmetrical data should have a skew near zero. Negative values of skew indicate data that are skewed left, whereas positive values of skew indicate data that are skewed right. By skewed left, we mean that the left tail is long relative to the right tail. Similarly to skewed left, skewed right means that the right tail is long relative to the left tail. CompassAA utilizes the following equation (9) for calculating skew:

$$\frac{n}{(n-1)(n-2)}\sum\left(\frac{x_1 - \bar{x}}{s}\right)^3 \qquad (9)$$

## Original document attributed to:

Sedorovich, A., Schuckman, K., O'Hara, C.G., ©2009 SIS: Accuracy Analyst Technical Reference - Statistical Methods, SISDOCS_MATAA_v1.0_TR002_060109

## Selected References

"Analytical Photogrammetry Instrumentation" [Online] <http://www.pharmacy.ferris.edu/faculty/burtchr/sure440/notes/Corrections_to_photo_coordinates.htm>

Congalton, R.G., Green K., 2009. Assessing the Accuracy of Remotely Sensed Data: Principles and Practices, CRC Press Taylor and Francis Group, Boca Raton, Florida, 183p.

Dana, P. H. "Global Positioning System Overview." [Online] September 1994. <http://www.colorado.edu/geography/gcraft/notes/gps/gps.html>.

Department of Geosciences at Idaho State University. "Introduction to Topographic Maps." [Online] 7 April 2008. <http://geology.isu.edu/geostac/Field_Exercise/topomaps/map_proj.htm>.

GIS Development. "Global Positioning System- An Overview" [Online] < http://www.gisdevelopment.net/technology/gps/techgp0038.htm>

National Geospatial-Intelligence Agency, cited 2009: Geolocation Accuracy Evaluations of OrbView-3, EROS-A, and SPOT-5 Imagery.

O'Hara, C.G., Cheriyadat, A., 2008. Accuracy Analyst: A Horizontal Accuracy Assessment Tool.

"Photogrammetric Surveys" [Online] < http://www.state.nj.us/transportation/eng/documents/survey/Chapter7.shtm>

Rose, C. M., 1991: Error Theory as Applied to Mapping, Charting, and Geodesy. DMA Tech. Rep. DMA TR 8400.1, 70 pp.

The Aerospace Corporation. "GPS Primer." [Online] 29 April 2005. ™ <http://www.aero.org/education/primers/gps/index.html>.

“The Figure of the Earth.” [Online]
<http://kartoweb.itc.nl/geometrics/Reference%20surfaces/body.htm>.

United States Bureau of the Budget, cited 2009: U.S. National Map Accuracy Standards. [Available online at http:// http://rockyweb.cr.usgs.gov/nmpstds/nmas.html]

World Health Organization, cited 2009: Guidelines for Data Collection in the Field using Global Positioning System (GPS) Technology. [Available online at http://www.who.int/entity/health_mapping/resources/GIS_guidelines_data_collection.pdf]