DOI: 10.2352/issn.2168-3204.2016.1.0.175

Rethinking Image Color Correction, Validation and Testing

Don Williams: Image Science Associates and Peter D. Burns: Burns Digital Imaging

Abstract

Digital image capture normally includes a color-correction step that transforms detector signals into corresponding image pixel values. For digital cameras and scanners, we usually base the color-correction operation on captured images of reference color charts. Measures of object color-capture are included in recent imaging guidelines for cultural heritage institutions. Several methods have been adopted as standard practice, with the aim of reducing image-capture variation. During the evaluation of the goodness of object-to-image color-encoding, there is normally a validation step. This involves comparing the original target colorimetry to that of the predicted colors and calculating summary color-difference metrics for the population of target samples.

While this is an instinctive and common approach we believe it needs to be revisited. The current summary statistics for evaluating color capture goodness can be misleading for the color-content at hand. Additionally, reporting color error measurements for the same colors that were used to develop the color-correction is effectively 'teaching to the test' when evaluating digital capture color performance. We discuss strategies for selecting validation colors based on generic and specific use cases along with examples.

Introduction

Color image capture includes a color-correction step that transforms detected color signals into corresponding pixel values. For digital cameras and scanners, we usually base the color-correction operation on captured images of reference color charts. However, all color image capture is based on information, and several assumptions, about the scene/object. For cultural heritage imaging, it has been proposed that color-correction should be based on test objects that reflect the spectra-color characteristics of the collection materials. References [1-4] address such collection-specific color correction and testing.

In this paper, we primarily address the evaluation of color capture, rather than methods for refining the color-image processing operations. As an example, Figure 1 shows two targets placed above a painting being photographed. Measures of the goodness of object color-capture based on such targets are included in recent imaging guidelines [5, 6] for cultural heritage institutions.

From a colorimetric description of the input reference color patches, and the corresponding unprocessed pixel values, we compute the color-correction parameters required for accurate color (-encoded) images. This signal transformation usually takes the form of either a custom or standard ICC profile to a color-space such as *sRGB* or *AdobeRGB*. We can cast the building of an ICC color profile as a statistical modeling operation, where the model takes on the form specified by the profile elements; look-up tables, color matrix, etc. One can consider the color profile as a code value-to-color dictionary.

In evaluating color image capture (including any ICC profile), there is normally a validation effort aimed at determining goodness of the color-correction operation. This involves comparing the target colorimetry to that predicted from the color profile-processed pixel values. Several visual color difference formulas are used for this. Figure 2 outlines the steps often used for this evaluation of color accuracy based on test chart patches. A commonly used measure is a computed as a CIELAB color-difference, $\Delta E^{\dagger}$, for each color patch. Summary ΔE statistical measures of the entire population of test target color patches are often used for reporting overall color-encoding performance. To date these have included simple average and maximum ΔE reporting. We propose that these naïve performance measures are reconsidered, and amended to provide greater specificity for varying collection color-content.

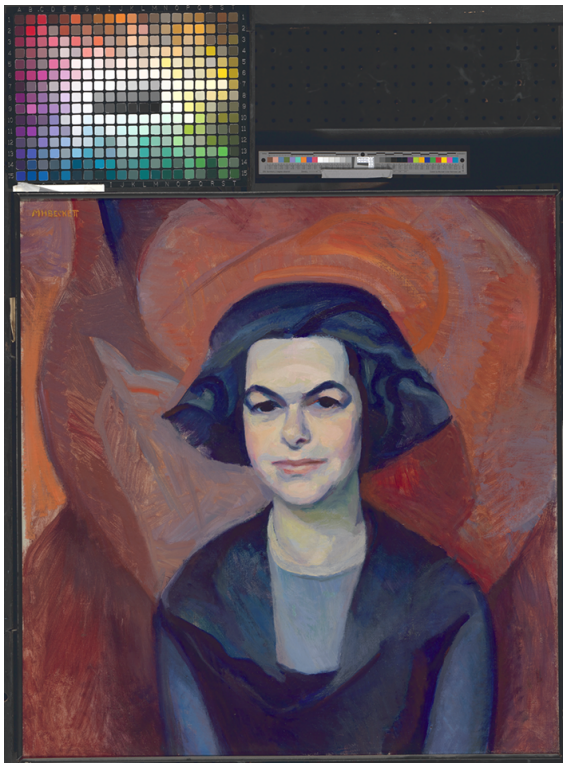

Figure 1: *Example of color test charts to control and evaluate image capture (courtesy of the US National Portrait Gallery, Washington, DC)*

[†] We use ΔE to indicate a general CIELAB distance such as ΔE_{ab}, ΔE_{94}, ΔE_{2000}.

DOI: 10.2352/issn.2168-3204.2016.1.0.175

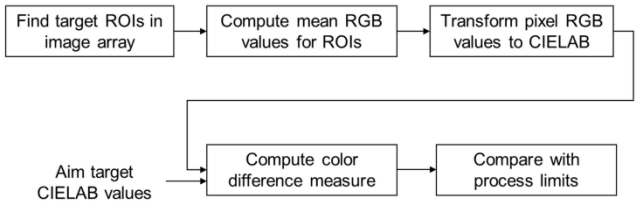

Figure 2: *Common procedure for evaluation of color-encoding accuracy, where the input is a captured image array*

It is also common to report color-encoding accuracy results using the same color-chart that was used to determine the color-correction parameters. However, when we do this we will usually under-estimate the residual color error.[†] We are effectively 'teaching to the test' when evaluating digital capture performance on the same colors by which the regression was performed. By definition, the reported error will be the minimum possible when using the calculated color profile model. This is in fact the role of regression (*i.e.* color-correction), to minimize the difference between a known standard and an estimate of that standard, consistent with the underlying statistical model selected for the color profile.

We investigate alternative reporting and validation approaches[7] for color error, where performance is judged against;

- more logical and useful summary statistical measures that are content specific
- an independent and different set of color patches.

Results for several strategies for selecting validation and summary colors metrics will be presented and discussed.

Summary Statistical Metrics

Consider the capture of the 300 patch [3] in the upper left of Fig. 1. We now follow a typical calibration and evaluation workflow to demonstrate our proposals for improved summary metrics and the specificity surrounding them.

- Using a capture of that 300 patch color target an ICC color profile was calculated using a color look up table (*cLut*) model.
- The ICC color profile was embedded into the image.
- The 300 patch target image was then evaluated against its colorimetric reference file for color-encoding error with respect to ΔE_{2000} honoring (using) the ICC profile.
- The individual color patch errors along with average, median, and maximum summary measures for all 300 patches are presented in Figure 3. This includes both pseudo-color and text based presentations of the errors.

Notice, in general, that a number of the higher error patches (yellow and red colored) are concentrated near the center set of 18 gray patches indicated by red box outline. These particular patches are of relatively high luminance with very low chroma components, considered *near-neutrals*.

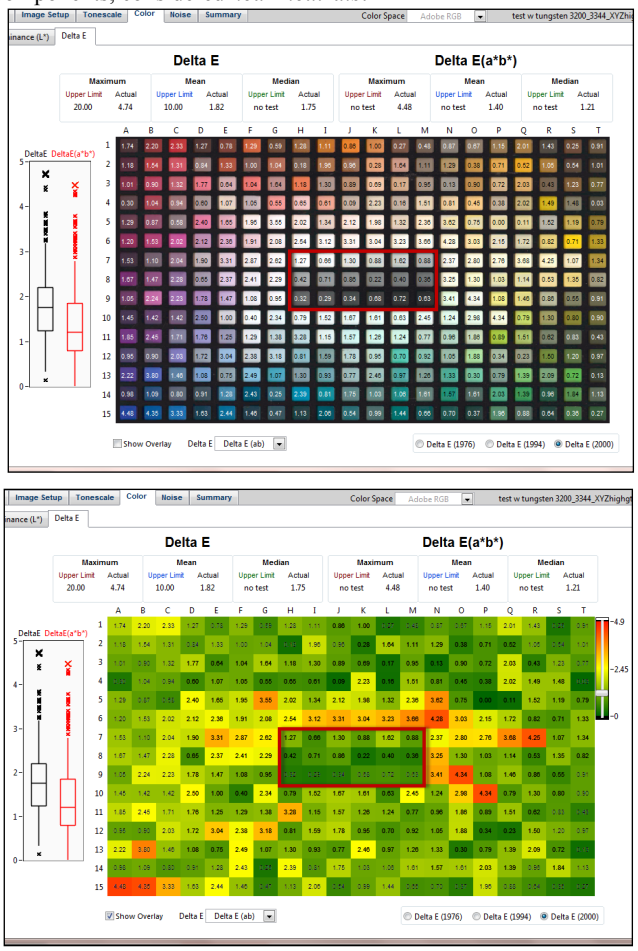

Figure 3: *Color-encoding error reports for the 300-patch target.*

In addition to reporting the average, maximum, and median ΔE_{2000} for the 300 patches a presentation of the distribution of color errors is provided in the box-whisker graphs on the left hand side of the panel. (See appendix for explanation of how to interpret these plots). These give a more thorough summary of the distribution of the errors. The top-bottom boundaries of the boxes are indicators of the quartile spread of the color errors while the 'X' markers are indicators of outlier patches. The spread metric is a precision indicator while the outliers can be symptomatic of either weaknesses in the profile model calculation or inaccurate reference data.

Another promising summary metric for evaluating the distribution of color errors is the cumulative percent point for a targets color patches. For instance, BasICColor Input color profiling software now reports ΔE population percentiles, typically 90%. Upper quartile metrics based on the median statistic are also used and will be demonstrated here. These approaches are much better than isolated ΔE_{max} values, since they provide context to the population of errors.

We propose that, at a minimum, a simple median metric be substituted for simple average since it is more resilient to outlier

[†] We use color *error* to mean any difference between ideal and observed. Some sources may be predicable due to system design or material characteristics; others may be stochastic in nature.

DOI: 10.2352/issn.2168-3204.2016.1.0.175

influences. If an average calculation is used there should be some accommodation or standard for outlier rejection. We suggest that the ΔE_{max} be abandoned altogether for the reasons stated above and that a cumulative percentile or quartile metric be adopted.

Color Content Specificity

It is instructive to compare the color content of the target patches specifically to the object being digitized. It would be logical to evaluate color-encoding goodness not by the target itself but rather by the extent to which the target represents the colors of the content under consideration. A simple approach is to exclude from the summary ΔE metric calculation, cited above, those target patches that are not representative of an object's color content.

We demonstrate this simply by eliminating from consideration all target patches that are not representative of the luminance data of the painting in Fig. 4. A simple histogram analysis on the luminance data of the object reveals that the object contains no luminance data above a count value of 200.

To truly evaluate the impact of the color profile on the object of interest, all patches above that value in Fig. 3 can be toggled off in the ΔE_{2000} calculation. This is illustrated graphically in Fig. 5 below. Note in particular that the first five highest gray patches within the red box are excluded. The box-whisker plots of Fig. 6 show the distribution differences graphically. Table 1 compares the summary statistics from the box-whisker plots with and without the high L^* values excluded. There is a significant decrease in the median color error as well as for the upper quartile value. This result is the focus of our proposal for more logical color specificity reports and auditing of the color-encoding errors.

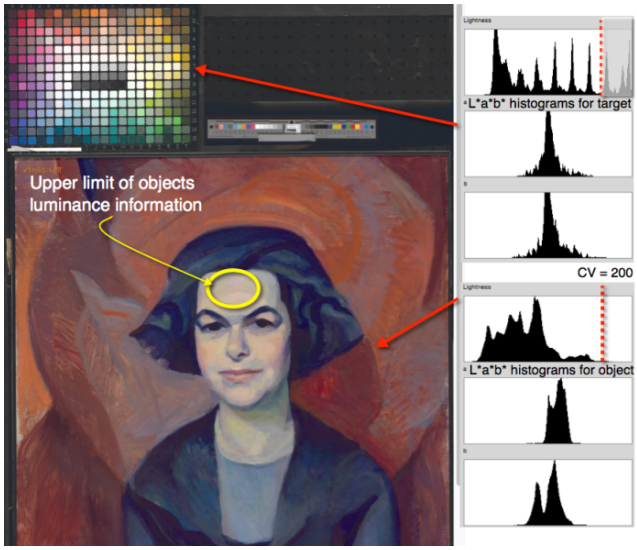

***Figure 4**: Luminance comparison between target and object.*

While the results presented here demonstrate a reduction in the effective color-encoding errors, this may not always be the case. It is just as likely that they may increase with object specific colors. Whatever the choice, this is a more logical and appropriate approach to performance reporting than the current method based on the target content alone. Targets can indeed be effective tools in assessing image capture performance but should not be considered image content themselves.

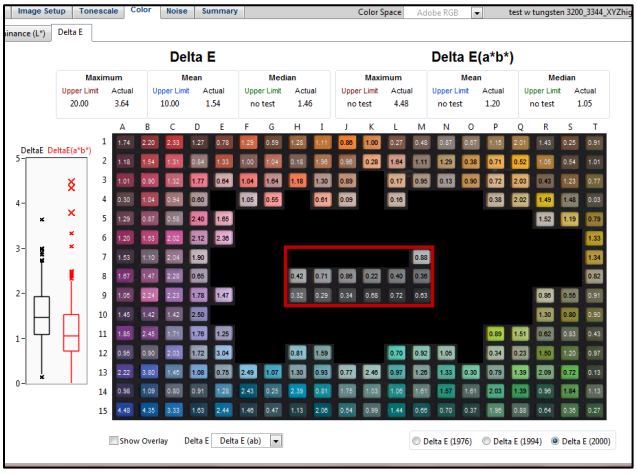

***Figure 5**: Excluded patches for object specific color error reporting*

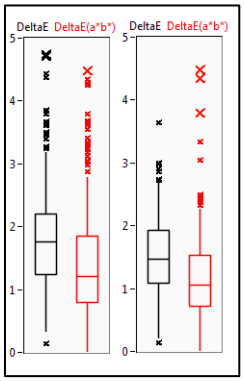

***Figure 6:** Box-Whisker plots with and w/o excluded L* values*

Table 1: Summary statistics for global and excluded patches

	ΔE_{2000}		$\Delta E_{ab-2000}$	
	all	L*< 200	all	L*< 200
Median	1.75	1.46	1.21	1.05
Upper quartile	2.24	1.90	1.85	1.50

Independent Color Testing

Validation Color Patch Selection Set

A popular way to evaluate the quality of this color calibration is to simply compare the translated color of each patch in Profile Connection Space (PCS), *L*a*b** or *XYZ*, to the measured reference color of the actual target. While this is an instinctive approach, it yields, by definition, an optimal residual color error for that model since the regression model is designed to minimize such errors. One is effectively 'teaching to the test' when

DOI: 10.2352/issn.2168-3204.2016.1.0.175

evaluating digital capture color performance using the same colors for which the color-correction was performed.

We suggest using a validation approach where the color performance is tested with an independent and different set of color patches. Borrowing from medical clinical trials, these can be thought of as control (calibration) and treatment (validation) groups. While color calibration- or profiling validation is not often discussed in the literature, it can provide valuable information regarding the quality of image capture, and the likelihood of color artifacts during normal operation of the image capture system.

We recognize that the strategy for selecting a validation set of colors is open to infinite opinions. Being reasonable and without focusing on building a 'killer' validation target, we restrict our patch selection for validation using a set of criteria already included in the popular ColorCheckerSG target (SG). They are,

- The same number of total patches
- Identical set of gray patches (61)
- Same number of chromatic patches (79)
- Same number of patches within L^*(10) slices
- Semi-Gloss surface
- Remained within the gamut of the existing SG

The differences between the two sets are,

- Different set of chromatic patches
- Select patches from the Natural Color System (NCS) index[5]

We will refer to the independent validation target as SG-X consistent with the criteria described above. Example images of the SG and SG-X targets are shown in Fig. 7. Details of the how these patches were selected are described in Ref 4.

Experimental

Consistent with common field practices, ICC color profiles were generated from the SG chart. We generated three different regression models using Rough Profiler, built on the open-source Argyll color-management system[8]. These models are labeled as,

1. *Lab cLut*, medium quality (*Lcm*)
2. Shaper + Matrix, medium quality (*SMm*)
3. *Lab cLut*, high quality (*Lch*)

The color profiles from these models were then embedded into the SG chart image from which the profiles were generated, in addition to the independent SGX validation chart image. It is the latter pairings that are of interest where the validation target (SGX) is assessed for color-encoding accuracy using a profile generated via the SG target. Results for several combinations of image target and ICC profile are presented in the results section that follows.

***Figure 7**: Comparison images of the SG (top) and SGX(bottom) targets*

Results

Table 2 summarizes the median color error *ΔE2000* results for the combination of different color profile models and target types.

Table 2: Summary ΔE2000 for SG generated profiles against indicated target type and regression model

ΔE_{2000}	Target type		Model
	SG	SGX	
median	2.18 (2.13)*	2.30 (2.34)*	Lcm
median	3.14	2.87	SMm
median	2.16	2.30	Lch

For the *Lcm* model alone we also normalized the ΔE_{2000} data by eliminating the common neutral values from the calculations. These data are shown in the parenthetical values of Table 2. This is also shown graphically in the pseudo-color illustrations of Figs. 8.

The median color-encoding error differences between the regression model types were mixed. As expected, there were increased median color errors of about 6% for the SG-SGX profile-target combination, but only for the more sophisticated *cLut* regression models. The difference was slightly higher (9%) for the neutral normalized data. There was actually a decrease in the median value for the *SMm* model. This decrease is unexplained.

However, greater insight can be gained from the pseudo color error graphics of Fig. 9. These map the color errors from low to high (green to red) of the *Lcm* profile comparison. Visually, there appears to be a greater color bias into the yellow and red colors for

DOI: 10.2352/issn.2168-3204.2016.1.0.175

the SGX_{Lcm} pairing than evident from the median summary values of the first row of Table 2. The box-whisker plots reconcile this by indicating a greater spread between the upper and lower quartiles for the SGX_{Lcm} pairing

There is a higher variance of ΔE_{2000} errors with the validation set than for the calibration set. So, while the overall accuracy (i.e. median) of the data for the SG-SGX is lower, albeit small, the variance of error is higher (i.e. lower precision) for this set. In combination the additive difference of lower accuracy and precision suggests a recommendation for performing independent calibration-validation tests, as outlined in this experiment.

This also indicates the need for better summary and specificity measures for evaluating color-encoding error for digital capture. We have chosen to use box-whisker plot because they can describe central tendency, distribution, and outlier data quickly in a simple graphic.

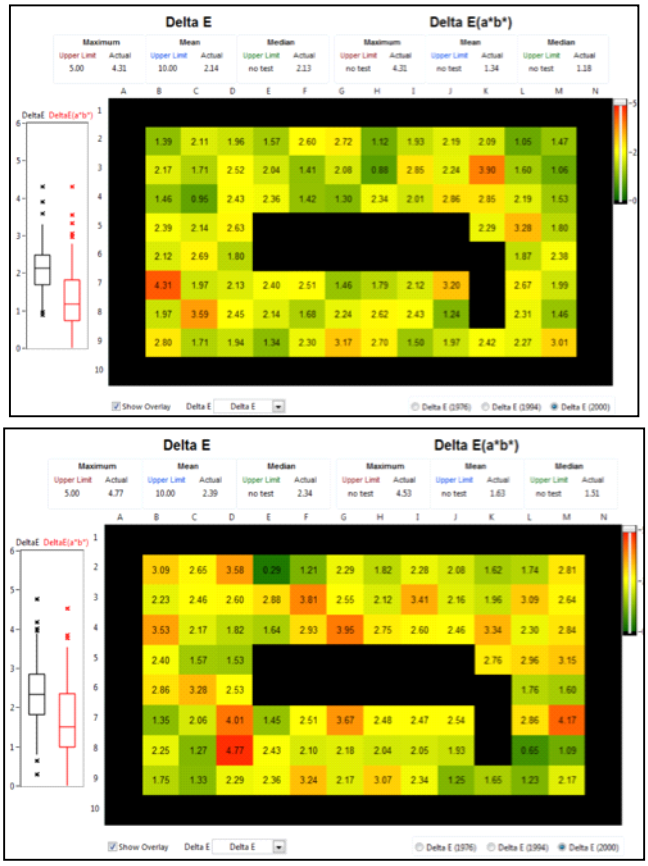

***Figure 8**:ΔE_{2000} map of SGLcm (top) and SGXLcm (bottom) target-profile combinations with neutrals toggled off*

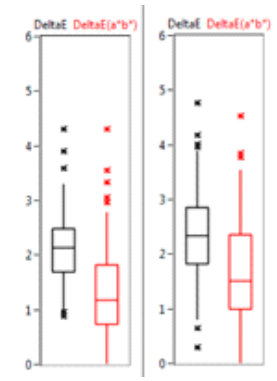

***Figure 9:** Box-whisker plots of ΔE_{2000} for SG-SG (left) and SG-SGX (right) profile/target combinations using the Lcm model*

Conclusions

We have demonstrated two different approaches for improved reporting of summary measures of color accuracy for digital capture. One uses existing target profiling techniques but focuses on the color content of the collection being captured. For example, as described in Ref. 4. The color target is viewed as a tool in enabling the measurement of color-encoding error, but may include non-representative samples. We believe these colors should be eliminated from consideration in reporting color-encoding errors, based on their importance.

Our second proposal for color error reporting is perhaps more important in evaluating the goodness of color calibration or profile generation. Instead of measuring and reporting residual error on the same set of colors used for calibration, we believe that an independent, but representative, set of colors be used for validation purposes.

For both proposals we suggest that the ΔE_{max} measure be eliminated as a summary for color accuracy reporting. Instead more robust, noise-resistant metrics like the median should be adopted. Also, measures that summarize the distribution of color errors should be considered. Box-whisker plots that use quartile statistics, or simple cumulative percentage values, are very good alternatives.

Acknowledgement

We have benefited from discussions with Prof. Roy Berns, Rochester Institute of Technology.

References

[1] G. Trumpy, Digital Reproduction of Small Gamut Object: A Profiling Procedure based on Custom Colour Targets, Proc. CGIV Conf., 143-147, IS&T, 2010

[2] D. Williams and P. D. Burns, Capturing the Color of Black and White, Proc. IS&T Archiving Conf., 96-100, IS&T, 2010

[3] R. S. Berns and M. I. Haddock., A Color Target for Museum Applications, Proc. Color Imaging Conf., 27-30, IS&T/SID, 2010

[4] D. Williams and P. D. Burns, Targeting for Important Color Content: Near Neutrals and Pastels, Proc. Archiving Conf., IS&T, 190-194, 2012

[5] FADGI Still Image Working Group, ed. T Rieger,, US Library of Congress http://www.digitizationguidelines.gov/

DOI: 10.2352/issn.2168-3204.2016.1.0.175

[6] H. van Dormolen, *Metamorfose Preservation Imaging Guidelines,* Image Quality version 1.0, Nat. Library of the Netherlands, 2012. https://www.metamorfoze.nl/sites/metamorfoze.nl/files/publicatie_documenten/Metamorfoze_Preservation_Imaging_Guidelines_1.0.pdf

[7] D. Williams and P. D. Burns, Color Correction Meets Blind Validation for Image Capture: Are We Teaching to the Test?, Proc. Electronic Imaging, IS&T, 2016 (in press)

[8] Argyll Color Management System, http://www.argyllcms.com/

Authors Biographies

Don Williams is founder of Image Science Associates, a digital imaging consulting and software group. Their work focuses on quantitative performance metrics for digital capture imaging devices, and imaging fidelity issues for the cultural heritage community. He has taught short courses for many years, contributes to several imaging standards activities, and is a member of the Advisory Board for the interagency US Federal Agencies Digitization Guidelines Initiative, FADGI.

Peter Burns is a consultant supporting digital imaging system and service development, and related intellectual property efforts. Previously he worked for Carestream Health, Eastman Kodak and Xerox Corp. He is a frequent conference speaker, and teaches courses on these subjects.

Appendix - interpreting box-whisker plots

The examples and wording here are largely taken from the National Instruments Labview manual on box-whisker plots. Only minor changes have been made. The box-whisker plots in this paper are taken from Labview software graphics.

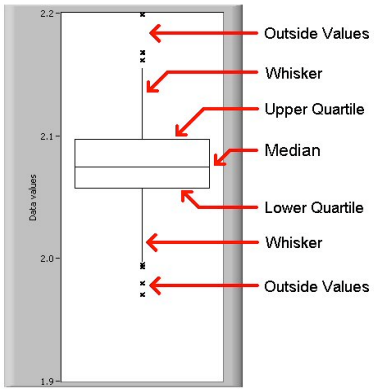

***Figure 10**: Example box-whisker plots*

The large, divided rectangle in the middle forms the box around which supplemental statistical features are derived. The upper and lower quartiles of the data set determine the size and location of this box. The line that divides the box horizontally through the middle represents the median of the data set. The top edge of the box is the value corresponding to the upper quartile of the data. The upper quartile is the median of the upper 50% of the data values, or the values greater than the global median. The bottom edge of the box is the value corresponding to the lower quartile of the data. The lower quartile is the median of the lower 50% of the data values, or the values less than the global median.

Vertical lines called whiskers extend from the middle of the top and bottom edges of the box. The whiskers are 1.5 times the inner quartile spread in length measured from the median. The inner quartile spread is the difference between the upper and lower quartiles of the data. The whiskers provide an arbitrary cutoff point to identify outlier values. Data points falling outside the whiskers but less than three times the length of the inner quartile spread are identified with small x's. Points beyond the whiskers are identified with large X's