Training Data Selection Study for Surface Colour Measurement Data Correlation

Thorsten Steder, M. Ronnier Luo, and Changjun Li; Department of Color Science, University of Leeds; Leeds LS2 9JT, UK

Abstract

In 1996 the National Physical Laboratory (NPL) in the UK conducted a project to determine the agreement of colorimetric measurements between different European laboratories. The results suggested significant measurement differences between instruments even though they were of similar design and built by the same manufacturer.

Modeling random and systematic spectrophotometric errors and the use of multiple regression analysis improved successfully the agreement between instruments since the early 1980s. This was in particular the case if they were regressed on each wavelength. Recently a new model based on band pass error was developed. It was said to outperform other models inherent of more errors in the equation. This paper shows that this was hardly the case.

Furthermore, it was evident that all models performed best when training and testing samples were the same. This showed clearly the dependency of the model on the physical properties of the samples used for training. Using other materials for testing resulted in just little improvement.

It was then of interest to determine which and how many samples were needed for training the model while maintaining a good performance. A method was found to reduce the number of training samples from a larger population of the Munsell Color Book. This resulted in a training set of 20 samples compared to 245 colour samples for modeling the correlation between two instruments with similar results.

Introduction

Accurate surface colour measurement plays an important role in industrial applications. For instance, in the colour industry, often production and quality control sites are located in different parts of the world. Communication can be done by electronically means (fax, email), or by sending physical samples for inspection. However, in both cases the samples will be measured to obtain reflectance data. This procedure should avoid miscommunication across sites, countries, and cultures.

To assure this, it is of vital importance that the measuring instruments across sites produce similar results for the same samples. Often different types and makes of measuring instruments were employed. This led to a project in 1996 conducted by the National Physical Laboratory (NPL) [1] in the UK in conjunction with 24 research laboratories across Europe being asked to measure the same samples for further considerations.

The results showed that the typical discrepancies between all laboratories were between 0.53 and 3.0 CIELAB units. Note that each manufacturer traces measurement results to one of the national bodies. The disagreements between national laboratories were included in the instrumental disagreement. Therefore, data correlation methods between different instruments can be used to obtain more consistent and better matches between them.

The earlier work in inter-instrumental agreement was given by Robertson [2]. This was further refined and extended by Berns and Petersen [3]. The mathematical model (Case 3 - wavelength dependent model) was then tested by Morovic *et al.* [4]. Following seven systematic errors were included: photometric zero, photometric linear scale, photometric non-linear scale, wavelength linear scale, wavelength non-linear scale quadratic, wavelength non-linear scale sine wave, and bandwidth. Morovic *et al.*'s model correlating one instruments' measurement to another is given in equation 1.

$$\mathbf{R_1(\lambda) = R_m(\lambda) + B_0X_0(\lambda) + B_1X_1(\lambda) + \ldots + B_6X_6(\lambda),} \text{ where} \quad \mathbf{(1)}$$

$X_0(\lambda) = 1$, $X_1(\lambda) = R_m(\lambda)$, $X_2(\lambda) = [100 - R_m(\lambda)]R_m(\lambda)$, $X_3(\lambda) = dR_m/d\lambda$, $X_4(\lambda) = w_1(\lambda)dR_m/d\lambda$, $X_5(\lambda) = w_2(\lambda)dR_m/d\lambda$, $X_6(\lambda) = d^2R_m/d\lambda^2$. The coefficients B_0, B_1, …, and B_6 vary with wavelength.

Recently, Chung *et al.* [5] developed a model based on band pass error. He claimed to have outperformed considerably the models of Morovic *et al.* [4] and Berns and Petersen [3]. The best results were obtained if the model was trained on the BCRA Series II matt tiles in specular excluded mode. Their model is given in equation 2.

$$\begin{aligned}
\mathbf{R'_{(400)}} &= \mathbf{n_{(400)}R_{(400)} + o_{(400)}R_{(410)} + p_{(400)}} \\
\mathbf{R'_{(\lambda)}} &= \mathbf{m_{(\lambda)}R_{(\lambda-10)} + n_{(\lambda)}R_{(\lambda)} + o_{(\lambda)}R_{(\lambda+10)} + p_{(\lambda)}} \quad \mathbf{(2)} \\
\mathbf{R'_{(700)}} &= \mathbf{m_{(700)}R_{(690)} + n_{(700)}R_{(700)} + p_{(700)}}
\end{aligned}$$

In order to correct the measurement of the test instrument ($\mathbf{R_{(\lambda)}}$ at a particular wavelength λ) and to correlate it with the measurement of the reference instrument ($\mathbf{R'_{(\lambda)}}$ at a particular wavelength λ) two neighbouring reflectance values (one from the right and one from the left of the wavelength scale) were required to predict $\mathbf{R'_{(\lambda)}}$. To improve this model it was extended by adding more reflectance values on either side. This model is called the 'Extended Chung model' as described in equation 3.

$$\begin{aligned}
\mathbf{R'_{(400)}} &= \mathbf{m_{(400)}R_{(400)} + n_{(400)}R_{(410)} + o_{(400)}R_{(420)} + p_{(400)}} \\
\mathbf{R'_{(410)}} &= \mathbf{m_{(410)}R_{(400)} + n_{(410)}R_{(410)} + o_{(410)}R_{(420)} + p_{(410)}R_{(430)} + q_{(410)}} \\
\mathbf{R'_{(\lambda)}} &= \mathbf{l_{(\lambda)}R_{(\lambda-2)} + m_{(\lambda)}R_{(\lambda-1)} + n_{(\lambda)}R_{(\lambda)} + o_{(\lambda)}R_{(\lambda+1)} + p_{(\lambda)}R_{(\lambda+2)} + q_{(\lambda)}} \quad \mathbf{(3)} \\
\mathbf{R'_{(690)}} &= \mathbf{m_{(690)}R_{(670)} + n_{(690)}R_{(680)} + o_{(690)}R_{(690)} + p_{(690)}R_{(700)} + q_{(690)}} \\
\mathbf{R'_{(700)}} &= \mathbf{m_{(700)}R_{(680)} + n_{(700)}R_{(690)} + o_{(700)}R_{(700)} + p_{(700)}}
\end{aligned}$$

In spectrophotometric measurements, standards such as the BCRA Series II tiles (including 12 tiles gloss or matt) or German EBUCAM matt colour samples (21 samples), were used for the purpose of training the model. These colours were selected and produced since they were most appropriate for investigating the errors of the instruments. However, if the testing colour samples were of different physical properties, compared to the training samples, the performance of the models decreased, considerably [6].

Since all models were developed with the aid of 'training samples', it was also of interest to determine how many samples were actually needed to train a model and, secondly, how to select them. For instance, the Munsell Color Book compromises well over 1600 samples. To measure all of them would be time and cost consuming. Hence, it was desirable to reduce the number of samples to be measured for training the data correlation models.

Experimental Design

The data sets selected were: 12 BCRA Series II gloss tiles, 12 BCRA Series II matt tiles, 21 German standard EBUCAM samples), 245 sub-sampled colour patches from the Munsell Color Book (semi gloss paint on paper) and 15 textile samples.

The samples were measured twice on each instrument; the mean reflectance values were taken for further considerations. Three spectrophotometer were used for the measurements of the samples: a Gretag Macbeth CE-7000A (diffuse 8°, 6'' sphere size), a Gretag Macbeth CE-2180 (diffuse 8°, smaller than 6'' sphere size), and the X-Rite 938 (portable 0°/45°). Measurements were taken in specular included (SCI) and excluded mode (SCE).

The measurement data obtained from the instrument CE-7000A were used as the reference data since they were closest to the reflectance data given by the NPL for the glossy tiles. The data from the other two instruments were then used to predict the values of the reference instrument.

Three models were tested: Morovic *et al.,* Chung *et al.*, and the extended Chung *et al*. model. They were tested on the data sets specified above.

Finally, the 245 Munsell book color samples were used in conjunction with the Extended Chung *et al.* model to reduce the amount of samples needed for training the model.

Results and Discussion of Correlation Models

Inter-instrumental agreement is the agreement between the measurement results of the same sample by two different instruments. All data sets were measured on each instrument and compared with each other. The relationship between the performance of the models (mean ΔE_{00}) and the materials used for training and testing can be seen in Table 1. The training data sets are listed in the left column for all three models whereas the training sets are listed in row 2. In general, the agreement before modeling between instruments of similar design (sphere instruments) varied from 0.2 (Munsell SCI) to 0.57 (Textiles SCE) ΔE_{00} units. The agreement between instruments of different geometry varied from 0.65 (EBUCAM SCE) and 5.19 (BCRA gloss SCI) ΔE_{00} units. This gave evidence that correlation is necessary between instruments even for those with similar design.

Table 1: Summary of Inter-instrumental agreement between instrument CE7000A and CE2180 (similar design and same manufacturer) for various models and data sets (mean ΔE_{00})

CE-7000A /CE-2180	***Testing***					
Datasets	***BC G SCI***	***BC G SCE***	***EBU SCI BC M SCI***	***EBU SCE BC M SCE***	***MUNS/ SCI***	***MUNS/SCE***
Original Difference	***0.21***	***0.20***	***0.29***	***0.28***	***0.20***	***0.28***
Training	*CIEDE2000*	*CIEDE2000*	*CIEDE2000*	*CIEDE2000*	*CIEDE2000*	*CIEDE2000*
Morovic *et al. (performance between various data sets in mean ΔE_{00} units)*						
BCRA G SCI	0.10	0.38	0.26	0.28	0.20	0.37
BCRA G SCE	0.20	0.11	0.23	0.18	0.25	0.28
BC/EBU SCI	0.25	0.22	0.09	0.15	0.34	0.35
BC/EBU SCE	0.31	0.20	0.17	0.11	0.37	0.36
MUNS/SCI	0.19	0.29	0.37	0.36	0.11	0.26
MUNS/SCE	0.19	0.24	0.24	0.28	0.20	0.23
Chung *et al (performance between various data sets in mean ΔE_{00} units)*						
BCRA G SCI	0.10	0.25	0.26	0.27	0.25	0.29
BCRA G SCE	0.20	0.12	0.24	0.18	0.30	0.28
BA/EBU SCI	0.26	0.29	0.13	0.16	0.36	0.38
BA/EBU SCE	0.30	0.20	0.18	0.11	0.38	0.39
MUNS/ SCI	0.21	0.29	0.37	0.36	0.12	0.25
MUNS/ SCE	0.22	0.29	0.27	0.35	0.19	0.26
Extended Chung *et al (performance between various data sets in mean ΔE_{00} units)*						
BCRA G SCI	0.09	0.39	0.31	0.37	0.30	0.43
BCRA G SCE	0.20	0.08	0.30	0.27	0.31	0.38
BC/EBU SCI	0.27	0.31	0.13	0.16	0.37	0.39
BC/EBU SCE	0.31	0.29	0.17	0.11	0.37	0.38
MUNS/ SCI	0.19	0.27	0.36	0.35	0.10	0.24
MUNS/ SCE	0.21	0.29	0.27	0.31	0.31	0.20

The highlighted 'green' fields are the mean inter-instrumental results in ΔE_{00} units for those cases in which training and testing were made on the same data sets. All other results were a combination between various data sets. It can be seen that the results, in general, were better if the training and testing sets were the same. Furthermore, it was evident that Chung *et al.*'s model was not able to outperform other models inherent of more modeled errors in the equation. This was also the case if the model was trained on the BCRA Series II matt samples; and it did not improve the agreement for textile samples, too. If trained and tested on textile samples the results became similar to other models. The extended Chung *et al.* model has slightly improved the results for all data sets.

Selection and Discussion of Training Samples

In digital imaging the characterization of a device from device dependent RGB values to device independent CIE XYX tristimulus values is often done by using charts. These charts include many colour samples (e.g. in the range of 24 up to 500) with known reflectance values. The mapping performance between these two data sets is mainly determined by the polynomial regression employed and the selection of colours used for training these models. Generally, it can be said that the higher the order of the polynomial was, and the larger the amount of training samples were, the better the results became.

Nevertheless, the method of mapping between RGB and XYZ tristimulus values can be seen as somehow similar to the process of correlating one instrument's measurement responses to another. In this case, regression analysis is also used and the performance is determined by the predicted errors (and associated coefficients) and the number of training samples used for the modeling process.

Furthermore, it was important to use samples with similar physical properties. For instance, if a model was trained on 12 BCRA gloss tiles (SCI) and tested on a sub-sampled selection of 245 Munsell book color samples (SCI) the result of inter-instrumental agreement decreased in performance from 0.2 ΔE_{00} to 0.3 ΔE_{00} units after modeling. But, if trained and tested on the same Munsell Book Colors the agreement between two instruments improved from 0.2 ΔE_{00} to 0.1 ΔE_{00} units. This showed clearly the variation of performance with the change in substrates and the number of samples employed.

Since the measurement of a large number of samples was cumbersome, it became desirable to find a method to reduce the number of colours to be measured without decreasing the performance of the correlation models. Hunt [7] suggested to find samples that represents a good average of the entire population; Rich [8] had chosen colour samples covering an average of at least 5 – 6 hues at 3 or 4 chroma and 4 to 5 lightness levels.

In this manner the Munsell Color Book was initially sub-sampled to a selection of 245 out of 1605 samples covering approximately a wide color gamut at different lightness and chroma values (5R, 10R, 5YR, 10YR, 5Y, 10Y, 5GY, 10GY, 5G, 10G, 5BG, 10BG, 5B, 10B, 5PB, 10PB, 5P, 10P, 5RP, 10RP). Also, a range of grey samples differing in lightness values were included.

Cheung and Westland [9] have described two methods for reducing the number of samples to be measured with similar results as obtained from a larger population. Their work was inspired by earlier work from Hardeberg [10]. The authors adapted this approach and the result was a selection of 24 samples that were able to outperform traditional methods for characterization of digital cameras (e.g. Gretag Macbeth ColorChecker DC). They introduced two algorithms for the use in CIELAB space. The best results were obtained using the 'maxminc' algorithm. The main idea was to select samples that were as different as possible and, secondly, the samples whose closest neighbor in the already selected samples were as far away as possible. Furthermore, it was of interest to select colour samples (a.) to cover a wide gamut and (b.) to have them distributed evenly around the colour space within a range of lightness levels.

Our method was different as such we followed our sub-sampled approach by reducing the amount of colour samples furthermore. Firstly, by drawing a straight line from the highest chroma value sample to the midpoint of the diagram for each of the initial 10 hues chosen for the sub-sampled selection. The samples that were scattered away from this line were removed one by one for each hue. This has reduced the 245 samples to 114.

Consequently more samples were removed. Also, in cases were they were scattered away; and when they were positioned close to the line but in-between the highest and lowest chroma value for each hue. This has finally reduced the number of samples for each hue to two.

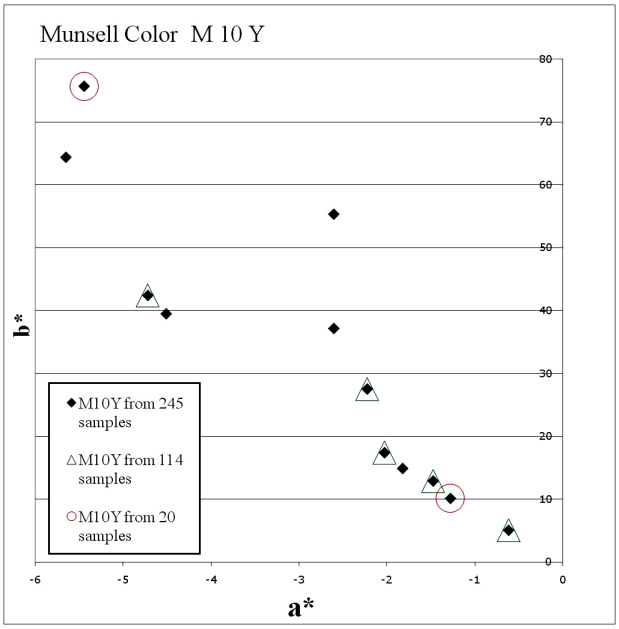

Figure 1: Reduction of colour samples for Munsell Colour 'M 10 Y'

An example of the result of the reduction method can be seen in Figure 2 (Munsell Book Color 'M 10 Y'). 12 colour samples were reduced, finally, to two samples. This corresponded to a selection of 20 samples for all 10 hues. The reduction method of the sub-sampled Munsell Color book has reduced the initial 245 samples to 114, 69, 58, 37, 20, 12, 10 and 6 samples for further considerations. The main difference between these two methods was the fact that it was not important to distribute the colour samples evenly spaced within the colour space over a wide range of lightness levels. Furthermore, at some stages the inclusion of neutral grey samples were omitted. From a selection of 20 samples downwards the hues were reduced to the five primary Munsell Book Colors.

Results

The results of the colour selection methods can be seen in Table 2. The performance was determined by the average ΔE_{00} unit agreement between two instruments (CE-7000A and X-Rite 938). All colour selection data sets were used to train the model (Extended Chung *et al*. model) while testing it on the full set of 245 colour samples.

Two main observations were made. First, the average colour difference between the two instruments was 0.27 ΔE_{00} units when the model was trained and tested on 245 samples. The training samples were then consequently reduced down to

10 samples. The best colour selection data set included 20 samples. The performance decreased when the number of samples fell below a size of 20 samples. The average colour difference between a model trained on 20 and tested on 245 samples was about 0.014 ΔE_{00} units.

Table 2: Difference of performance in mean ΔE_{00} units between training sets of 245, 69, 114, 58, 20, 37, 30, 12, and 10 samples and a testing set of 245 samples.

Extended Chung et al. – REF CE7000A - X-Rite 938 / Best Manual 20 Samples					
Col. Diff.	CIEDE2000	CIEDE2000	CIEDE2000	CIEDE2000	CIEDE2000
Samples	245	69	114	58	20
Average	0.27	0.28	0.28	0.29	0.29
MAX	2.95	2.98	3.06	3.07	2.98
MIN	0.02	0.03	0.02	0.04	0.03
MEDIAN	0.20	0.21	0.21	0.21	0.23
DIFF245	0.00	0.01	0.01	0.013	0.014
Samples	245	37	30	12	10
Average	0.27	0.29	0.29	0.44	0.45
MAX	2.95	3.0	3.0	5.8	3.0
MIN	0.02	0.02	0.02	0.02	0.03
MEDIAN	0.20	0.22	0.23	0.34	0.38
DIFF245	0.00	0.02	0.02	0.17	0.17

The colour selection process was then re-defined (according to 20 samples) such as only high chroma colour values were further reduced (e.g. from 80 to 60 and 40), or only low chroma values were increased (e.g. from 10 to 20 and 30). It was of interest to determine how this increase or decrease of the chroma values on either side of the scale would change the performance of the model. However, the results suggested no better performance in either ways.

For a comparison, the method used by Cheung and Westland [7] was implemented in Matlab with the aim to obtain an optimum colour sample set of 20 with the following constraints: The method was used four times only differing in the approach how to select the first and which sample for starting the computational procedure. Two different high chroma (HC1, HC2) and two different low chroma values (LC1, LC2) were randomly chosen from the reference set. The best selection proposed by this method for finding an optimum colour data set of 20 samples from a larger population of 245 Munsell Color Book samples resulted in an average inter-instrumental agreement of 0.36 ΔE_{00} units compared to 0.29 ΔE_{00} (manual mode selection of samples as proposed by the authors of this paper). The results can be seen in Table 3.

Table 3: Difference of performance of 'maxminc' algorithm for 20 samples in mean ΔE_{00} (first sample selection either of high 'HC' or low chroma 'LC')

Extended Chung et al. – REF CE7000A - X-Rite 938 / MAXMINC COLOUR SELECTION BEST OF 20 SAMPLES					
Col. Diff.	CIEDE2000	CIEDE2000	CIEDE2000	CIEDE2000	CIEDE2000
Samples	245	HC1	HC2	LC1	LC2
Average	0.27	0.38	0.42	0.40	0.36
MAX	2.95	3.02	3.09	3.13	3.03
MIN	0.02	0.03	0.01	0.02	0.02
MEDIAN	0.20	0.27	0.30	0.26	0.24
DIFF245	0.00	0.11	0.15	0.13	0.09

Conclusion

Chung *et al.*'s model was improved by adding more terms into the equation. Then, the model performed similar or better than Morovic *et al.*'s model. The initial Chung *et al.* model was not, as claimed, considerably better than Morovic *et al.*'s model. Neither in the case of being trained on matt samples in specular excluded mode, nor in conjunction with other data sets. Secondly, it was evident that the performance of the models was, indeed, material dependent. This agreed well with findings made in earlier studies. Furthermore, it was possible to reduce a sub-sampled selection of the Munsell Color Book to 20 training samples without decreasing the overall performance of the data correlation model, significantly. It was evident that the performance of the selection method was sensitive to low chroma values. The values around 10 were most appropriate. Furthermore, the hues (primary, secondary and a combination of them) should be uniformly distributed throughout the colour space.

References

[1] National Physical Laboratory (2006). Measurement Good Practice Guide No. 96: Surface Color Measurements, Teddington: NPL, UK

[2] Robertson, A.R., Diagnostic Performance Evaluation of Spectrophotometers. In. Advances in Standards and Methodology in Spectrophotometry, Elsvier, Amsterdam, 1986

[3] Berns, R.S. and K.H. Petersen, Empirical Modeling of Systematic Spectrophotometric Errors, Col. Res. Appl., **13**, (4), 08/1988, pp. 243 – 256

[4] Morovic, P., H. XU, and M.R. Luo, Inter-Comparison of Colour Measuring Instruments, Colour Image Institute, University of Derby, 1999

[5] Chung, Y.S., J.H. Xin, and K.M. Sin, Improvement of inter-instrumental agreement for reflectance spectrophotometers, Color. Technol., **120**, (2004), pp. 284 - 292

[6] Xun, Li., Wei Ji, Chanjun Li, Guihua Cui, and M.R. Luo, Comparison Study of the Surface Colour Measurement Data Correlation, 10th Congress of the International Colour Association, 8-13 May 2005, Granada Spain, pp. 725-728

[7] Hunt, R.W.G. (1998), Measuring Colour. Third Edition, Kingston upon Thames: Fountain Press, England

[8] Rich, D. C. and D. Martin, Graphic Technology – Improved model for improving the inter-instrumental agreement of spectrocolorimeter, NPES 2004

[9] Cheung, V. and S. Westland, Methods of Optimal Color Selection, Journal of Imaging Science and Technology, 50, **(5)**, (2006), pp. 481 - 488

[10] Hardeberg, J.Y, Acquisition and reproduction of colour images: Colorimetric and multispectral approaches, PhD Thesis, Ecole Nationale Superieure des Telecommunications, France, 1999

Author Biography

Thorsten Steder is a qualified Imaging Scientist (IQS, ARPS) and received his first class BSc (Hons) in Digital and Photographic Imaging from the University of Westminster, London, UK in 2006 and his MSc in Colour and Imaging Science from the University of Leeds, Leeds, UK in 2007. Since then he is a PhD student under the guidance of Prof. M. R. Luo. His interests are related to digital imaging devices, digital image processing, colour measurement, and colour difference formulae. This report is the summation of the results of his MSc project (author's email address: ccd6ts@leeds.ac.uk).