Comparison of Metrics: Discrimination Power of Pearson's Linear Correlation, RMSE and Outlier Ratio

Greg Cermak

Editor's note: This article by former ILG Co-Chair Greg W. Cermak appeared in the VQEG reflector in June, 2008 under the title "Comparison of Metrics: VQEG Multimedia Data." The article is reprinted with permission from Greg W. Cermak. Bracketed text and footnotes indicate clarifications by the editor.

The graphs and tables below show three things:

- [Pearson's linear] correlation, RMSE,[11] and outlier ratio all measure essentially the same thing.
- RMSE is better at discriminating between models.
- The advantage of RMSE over correlation increases as the number of video samples decreases, and vice versa.

These conclusions were also true in FRTV2.[12]

This note is organized into three parts. Part 1 shows the interrelationship of the metrics correlation (Pearson's R), RMSE, and the outlier ratio. Part 2 shows the performance of the metrics correlation, RMSE, and outlier ratio for the VQEG MM[13] data set. Part 3 shows the actual performance of R and RMSE for the FRTV2 and MM data sets and for hypothetical data from an experiment with 20 PVSs.[14]

RMSE is better than Pearson's linear correlation and outlier ratio at discriminating between objective video quality models.

[11] Root mean square error, Ed.
[12] VQEG's full reference television validation test, phase II, Ed.
[13] VQEG's multimedia validation test, phase I, Ed.
[14] Processed video sequence, Ed.

Part 1. Explanation

Each of the plots below is based on the FR metrics[15] for the 13-14 tests across 5 proponents; therefore 65-70 data points per plot. The metrics are highly correlated with each other.

Below the plots, for each resolution, is the output of a Principal Components factor analysis on the same data. The highlighted number labeled "proportion" is the proportion of variance in the 3 metrics across the 13-14 tests and 5 models that is accounted for by a single factor. That proportion of variance (an R2 measure) is always around 0.9, and the proportion accounted by any other factor is tiny. That is, each of the metrics is measuring essentially a single underlying factor, although in slightly different ways.

Following the graphs and factor analyses (Part 1) are the results of doing significance tests comparing each model to the best-performing model according to each type of metric, for each resolution (Part 2). These results are presented as tables of 1's and 0's. A '1' means that a model is tied with the top-performing model in the sense that it is not statistically significantly different. The more 1's in a table, the more ties. The more ties, the poorer the discrimination of the metric. Counting up the 1's, RMSE outperforms both correlation and outlier ratio in discriminating between models.

This analysis was critical in VQEG's decision to use RMSE to measure significant differences between objective video quality models in the HDTV validation test.

Correlation and outlier ratio have their advantages. Correlation is good for a simple summary of results. Outlier ratio is good for diagnosing model performance in order to improve the model's performance. When it comes to distinguishing between models, RMSE does the best job.

[15] Full reference metrics, Ed.

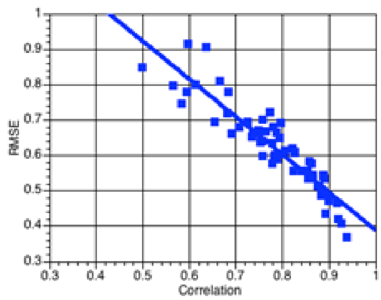

Figure 1. VGA data, RMSE plotted against Pearson's R. Linear fit R2 = 0.854, R = 0.92

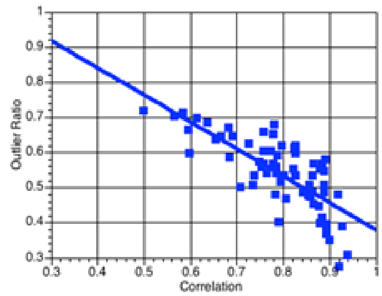

Figure 2. VGA data, Outlier Ratio plotted against Pearson's R. Linear fit R2 = 0.577, R = 0.760

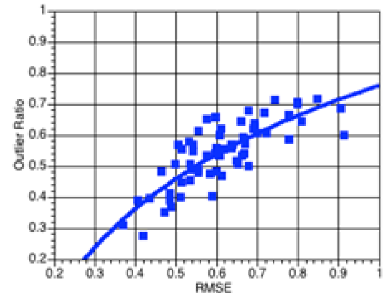

Figure 3. VGA data, Outlier Ratio plotted against RMSE. Log fit R2 = 0.653, R = 0.808

Principal Components analysis of the three metrics for VGA 1
10:04 Monday, May 5, 2008

The PRINCOMP Procedure

Observations	65
Variables	3

Simple Statistics

	corr	rmse	outratio
Mean	0.7901538462	0.6132461538	0.5412153846
StD	0.1000312812	0.1156449349	0.1012611865

Correlation Matrix

	corr	rmse	outratio
corr	1.0000	-.9240	-.7595
rmse	-.9240	1.0000	0.7858
outratio	-.7595	0.7858	1.0000

Eigenvalues of the Correlation Matrix

	Eigenvalue	Difference	Proportion	Cumulative
1	2.64838753	2.37151340	0.8828	0.8828
2	0.27687412	0.20213577	0.0923	0.9751
3	0.07473835		0.0249	1.0000

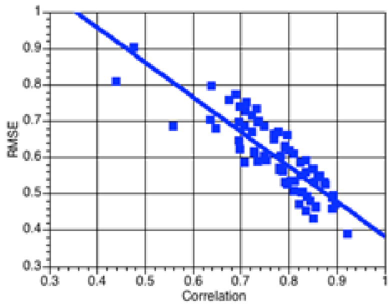

Figure 4. CIF data, RMSE plotted against Pearson's R. Linear fit R2 = 0.721, R = 0.849

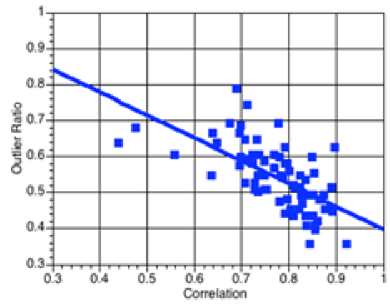

Figure 5. CIF data, Outlier Ratio plotted against Pearson's R. Linear fit R2 = 0.405, R = 0.636

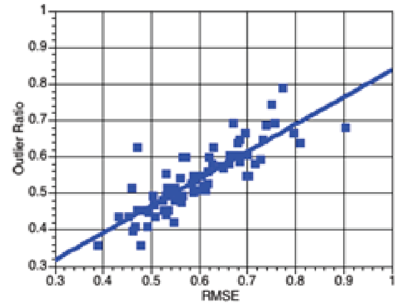

Figure 6. CIF data, Outlier Ratio plotted against RMSE. Linear fit R2 = 0.707, R = 0.841

Principal Components analysis of the three metrics for CIF 1
10:26 Monday, May 5, 2008

The PRINCOMP Procedure

Observations	70
Variables	3

Simple Statistics

	corr	rmse	outratio
Mean	0.7722857143	0.6013857143	0.5422428571
StD	0.0907168142	0.1022225893	0.0907040794

Correlation Matrix

	corr	rmse	outratio
corr	1.0000	-.8494	-.6362
rmse	-.8494	1.0000	0.8382
outratio	-.6362	0.8382	1.0000

Eigenvalues of the Correlation Matrix

	Eigenvalue	Difference	Proportion	Cumulative
1	2.55313907	2.18923473	0.8510	0.8510
2	0.36390434	0.28094776	0.1213	0.9723
3	0.08295658		0.0277	1.0000

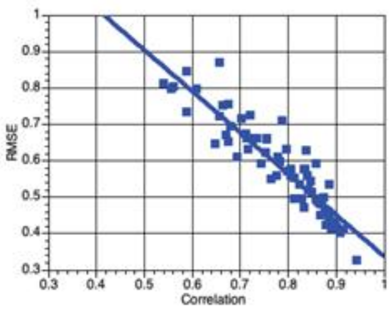

Figure 7. QCIF data, RMSE plotted against Pearson's R. Linear fit R2 = 0.853, R = 0.924

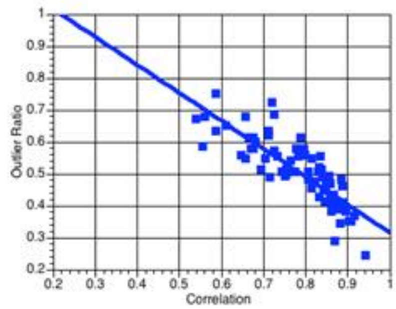

Figure 8. QCIF data, Outlier Ratio plotted against Pearson's R. Linear fit R2 = 0.702, R = 0.838

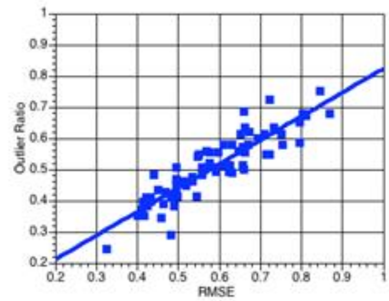

Figure 9. QCIF data, Outlier ratio plotted against RMSE. Linear fit R2 = 0.805, R = 0.897

Principal Components analysis of the three metrics for QCIF 1
10:48 Monday, May 5, 2008

The PRINCOMP Procedure

Observations	70
Variables	3

Simple Statistics

	corr	rmse	outratio
Mean	0.7816857143	0.5848142857	0.5069142857
StD	0.0997379963	0.1224948733	0.1043994201

Correlation Matrix

	corr	rmse	outratio
corr	1.0000	-.9235	-.8378
rmse	-.9235	1.0000	0.8972
outratio	-.8378	0.8972	1.0000

Eigenvalues of the Correlation Matrix

	Eigenvalue	Difference	Proportion	Cumulative
1	2.77288905	2.60847722	0.9243	0.9243
2	0.16441183	0.10171272	0.0548	0.9791
3	0.06269911		0.0209	1.0000

Part 2. Performance of the metrics Correlation, RMSE, and Outlier Ratio for the VQEG MM data set[16]

Editor's note: Rows contain objective video quality models. Columns contain subjective video quality datasets (e.g., V01, V02). The table title indicates the metric used to calculate statistical equivalence: Pearson linear correlation, RMSE, or outlier ratio.

Table 1. VGA data, correlation metric. FR Models

	V01	V02	V03	V04	V05	V06	V07	V08	V09	V10	V11	V12	V13	Total
Psy_FR	1	0	1	1	0	1	1	1	1	1	1	1	1	**11**
Opt_FR	0	1	1	1	1	1	1	0	1	0	1	1	1	**10**
Yon_FR	1	0	0	1	1	1	1	1	1	1	1	0	1	**10**
NTT_FR	1	0	1	1	1	1	0	1	0	0	0	1	1	**8**
PSNR DMOS	1	0	1	1	0	0	0	0	0	0	0	0	0	**3**

Table 2. VGA data, RMSE metric. FR Models

	V01	V02	V03	V04	V05	V06	V07	V08	V09	V10	V11	V12	V13	Total
Psy_FR	1	0	1	1	0	1	0	1	1	1	1	1	1	**10**
Opt_FR	0	1	1	1	1	1	1	0	1	0	1	0	0	**8**
Yon_FR	0	0	0	1	1	1	1	1	0	1	0	0	0	**6**
NTT_FR	1	0	1	1	0	0	0	1	0	0	0	0	0	**4**
PSNR DMOS	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**

[16] To assist in the readability of the tables in this reprint, (1) label "Total=" was replaced with "Total", (2) label "PSNR_DMOS" was replaced with "PSNR DMOS" and (3) the tables were transposed. As a consequence of the transposition, the label in the upper-left box ("Test") became incorrect and was omitted. See section 9 of the Multimedia Phase I ILG Data Analysis for these tables in their original format, Ed.

Table 3. VGA data, outlier ratio metric. FR Models

	V01	V02	V03	V04	V05	V06	V07	V08	V09	V10	V11	V12	V13	Total
Psy_FR	1	0	1	1	1	1	1	1	1	1	1	1	1	**12**
Opt_FR	1	1	1	0	1	1	1	0	1	1	1	1	1	**11**
Yon_FR	1	0	0	1	0	1	1	1	1	1	0	0	1	**8**
NTT_FR	1	0	1	1	1	0	1	1	1	0	0	1	1	**9**
PSNR DMOS	1	0	1	1	0	0	0	0	1	0	0	0	0	**4**

Table 4. CIF data, correlation metric. FR Models

	C01	C02	C03	C04	C05	C06	C07	C08	C09	C10	C11	C12	C13	C14	Total
Psy_FR	1	1	1	1	1	1	1	1	1	1	1	1	1	1	**14**
Opt_FR	1	1	0	1	1	1	1	1	1	1	1	1	1	1	**13**
Yon_FR	1	0	0	1	1	1	1	1	1	1	1	1	0	0	**10**
NTT_FR	0	1	1	1	1	1	1	0	0	1	0	1	0	0	**8**
PSNR DMOS	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**

Table 5. CIF data, RMSE metric. FR Models

	C01	C02	C03	C04	C05	C06	C07	C08	C09	C10	C11	C12	C13	C14	Total
Psy_FR	1	1	1	1	1	1	1	1	1	1	1	1	1	0	**13**
Opt_FR	1	1	0	1	0	1	1	1	1	0	1	1	0	1	**10**
Yon_FR	1	0	0	0	1	1	1	1	1	1	1	1	0	0	**9**
NTT_FR	0	1	1	1	0	1	1	0	0	0	0	1	0	0	**6**
PSNR DMOS	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**

Table 6. CIF data, outlier ratio metric. FR Models

	C01	C02	C03	C04	C05	C06	C07	C08	C09	C10	C11	C12	C13	C14	Total
Psy_FR	1	1	1	1	1	1	1	0	1	1	1	1	1	0	**12**
Opt_FR	1	1	0	1	1	1	1	1	1	1	1	1	1	1	**13**
Yon_FR	1	0	1	1	1	1	1	1	1	1	1	1	0	0	**11**
NTT_FR	1	1	1	1	0	1	1	0	1	1	0	1	1	0	**10**
PSNR DMOS	0	0	0	0	0	0	0	0	0	0	0	1	0	0	**1**

Table 7. QCIF data, correlation metric. FR Models

	C01	C02	C03	C04	C05	C06	C07	C08	C09	C10	C11	C12	C13	C14	Total
Psy_FR	1	1	1	1	1	1	1	1	1	0	1	1	1	0	**12**
Opt_FR	0	0	1	1	1	0	1	1	1	1	1	1	1	1	**11**
Yon_FR	1	0	0	0	1	0	0	1	0	0	1	0	0	0	**4**
NTT_FR	1	1	1	1	1	0	0	1	0	0	1	0	1	1	**9**
PSNR DMOS	0	0	1	0	0	0	0	0	0	0	0	0	0	0	**1**

Table 8. QCIF data, RMSE metric. FR Models

	Q01	Q02	Q03	Q04	Q05	Q06	Q07	Q08	Q09	Q10	Q11	Q12	Q13	Q14	Total
Psy_FR	1	1	1	0	1	1	1	1	1	0	1	1	1	0	**11**
Opt_FR	0	0	1	1	1	0	0	1	1	1	1	1	1	1	**10**
Yon_FR	0	0	0	0	0	0	0	1	0	0	1	0	0	0	**2**
NTT_FR	1	0	1	1	0	0	0	1	0	0	1	0	1	1	**7**
PSNR DMOS	0	0	1	0	0	0	0	0	0	0	0	0	0	0	**1**

Table 9. QCIF data, outlier ratio metric. FR Models

	Q01	Q02	Q03	Q04	Q05	Q06	Q07	Q08	Q09	Q10	Q11	Q12	Q13	Q14	Total
Psy_FR	1	1	0	1	1	1	1	1	1	1	1	1	1	0	**12**
Opt_FR	0	1	1	1	1	0	1	1	1	1	0	1	1	1	**11**
Yon_FR	1	1	0	1	1	1	1	1	0	0	1	0	0	0	**8**
NTT_FR	1	1	1	1	1	0	0	1	0	1	1	0	1	1	**10**
PSNR DMOS	0	0	1	0	0	0	0	0	0	1	1	0	0	1	**4**

Part 3. Comparison of R and RMSE for data sets of different sizes

(Thanks to Steve Wolf for a close reading and suggestions about this section.)

Consider the data from FRTV2. In FRTV2 there were two experiments, one for 525-line video and one for 625-line video. From FRTV2 we have, for six models that were in any kind of contention, correlation (Pearson's R) and RMSE scaled to a 5-point scale:

Table 10. FRTV2 data for 525 and 625 experiments, correlation and RMSE metrics.

R, 525 data	RMSE, 525 data	R, 625 data	RMSE, 625 data
0.937	0.37	0.898	0.395
0.935	0.375	0.886	0.415
0.856	0.55	0.884	0.42
0.836	0.585	0.87	0.445
0.756	0.695	0.779	0.565
0.682	0.775	0.703	0.64

In FRTV2, there were 64 PVSs in each experiment. In the MM experiments there were in excess of 150. For purposes of the following analyses, we consider experiments with 150, 64, and (hypothetically) 20 PVSs.

First, the critical difference in R required to declare two models different is given in sections 8.4.1 and 8.5.1 of the MM Final Report draft 1.4.1. The R to Z transform is applied, then the critical difference in Z-scores is computed; this critical Z depends on R and N, the number of PVSs in the test. Using a handy spreadsheet designed by Jamie DeCoster & Anne-Marie Leistico, we can determine that if R = 0.85 and N=150, then the critical R difference = 0.08. In Table 11 below, we also compute the critical R difference for R = 0.85 and N = 64 (the number of PVSs in FRTV2) and for N = 20.

The corresponding critical "RMSE difference" is actually a ratio of mean squared errors (MSEs) for any two models being compared. Given N, the critical F ratio is available from published tables or can be calculated in spreadsheets. We use critical F at the 95% confidence level for N = 150, 64, and 20.

Next we determine the corresponding RMSE's. We have empirical relationships between RMSE and R in Table 10 and in Figures 1, 4, and 7 above. Since there is not a single, unique relationship in our empirical data, we do computations for three different RMSE-R relationships given below (the one for VGA is very similar to the ones for CIF and QCIF):

- VGA: RMSE = -1.07*R + 1.46
- FRTV2 525: RMSE = -1.62*R + 1.91
- FRTV2 625: RMSE = -1.26*R + 1.53

Using these empirical relationships, and assuming that the target range of R's that are of interest is around 0.85, plus or minus some, we go through the following steps. These steps are based on being able to calculate critical R differences based on the Z-transform and known relations between N and Z; transforming from R to RMSE given the empirical relations above; calculating critical RMSE's from F-tables (based on corresponding MSEs and N); and transforming back to the familiar R scale using the empirical relations above. We then can compare critical differences in the data required for significance using R and RMSE. The steps:

1. For a given N (column 1 in Table 11), calculate the critical R difference (column 2 in Table 11). I used the spreadsheet by DeCoster & Leistico; in this example it is 0.08.
2. Using one of the empirical relationships above, find the corresponding RMSE. In the case of both VGA and FRTV2 525 it turned out to be 0.550.
3. Square the RMSE to find MSE.
4. For the given N, find the critical F value (for 95% confidence). For N = 150, that turns out to be 1.31 (column 3 in Table 11).
5. Find the critical MSE for the second model; in this example it is ((0.550)**2)*1.31 = 0.396.
6. Convert back to RMSE by taking the square root; in this example, it is 0.630 (column 4 in Table 11).
7. Find the corresponding R value using the empirical relationship above (for the VGA example given, this would be (0.630 – 1.46) / (-1.07) = 0.776.
8. Take the difference between the starting R (0.85) and the critical R (0.776); in the VGA example it is approximately 0.07 (column 6 in Table 11).

Following these steps, we get Table 11.

Table 11. Differences in Pearson's R required for statistical significance for three values of N, the number of PVSs, and corresponding RMSE differences (scaled in terms of R).

1	2	3	4	5	6	7	8
N	**R diff**	**F (.95)**	**Critical RMSE for VGA and FRTV2 525**	**Critical RMSE for FRTV2 625**	**Estimated R diff for VGA**	**Estimated R diff for FRTV2 525**	**Estimated R diff for FRTV2 625**
150	0.08	1.31	0.63	0.525	0.07	0.06	0.05
64	0.14	1.51	0.676	0.564	0.12	0.09	0.08
20	0.32	2.12	0.801	0.669	0.23	0.17	0.17

Table 11 shows that as N gets smaller, the critical R difference (column 2) and the corresponding critical RMSE (columns 4 and 5) both get larger, as we expect. However, the difference in sensitivity between R and RMSE also gets larger as the sample size decreases (compare column 2 with columns 6, 7, 8). Or, the other way around, if N gets very large, then the sensitivity of R and of RMSE probably converge. Also, we first noticed the advantage of RMSE in FRTV2 where the N was smaller than the recent MM project, and the consequent advantage in sensitivity for RMSE was more obvious.

Theoretical

Clearly, RMSE and R both depend on N and on the empirical distribution of discrepancies between model predictions and the observed MOS or DMOS scores[17] (called Perror in the MM Final Report). Presumably, one could write out the relationships between R and N and Perror, and between RMSE and N and Perror. Then it might be obvious when the critical R difference and the critical RMSE difference should differ from each other. I have not tried this exercise yet. Also, the empirical relationships between RMSE and R given above

[17] Mean opinion score (MOS) and differential mean opinion score (DMOS), Ed.

are certainly just estimates of some theoretically "true" relationship. Steve Wolf and I have made different guesses about what this relationship might be, but we are not quite ready to say what those guesses are.

Reference

"Draft final report from the video quality experts group on the validation of objective models of multimedia quality assessment, phase I," Version 1.4.1 April 15, 2008. ©2008 VQEG.

Greg W. Cermak performed the ILG's official data analysis for the VQEG Full Reference Phase I and II validation tests, the VQEG Multimedia validation test, and the VQEG RRNR-TV validation test. Greg was a Co-Chair of the VQEG Independent Lab Group (ILG) until he retired from Verizon in 2010.