Bulletin of Electrical Engineering and Informatics
Vol. 7, No. 1, March 2018, pp. 63~69
ISSN: 2302-9285, DOI: 10.11591/eei.v7i1.902

On Randomness of Compressed Data Using Non-parametric Randomness Tests

Kamal A. Al-Khayyat[1]*, Imad F. Al-Shaikhli[2], V. Vijayakuumar[3]
[1,2]Department of Computer Science, Islamic International University Malaysia, IIUM, Kuala Lumpur, Malaysia
[3]School of Computing Science & Engineering, Vit University, India

Article Info

Article history:

Received Nov 13, 2017
Revised Jan 14, 2018
Accepted Jan 28, 2018

Keywords:

Compression
Lossless
Non-parametric
Randomness

ABSTRACT

Four randomness tests were used to test the outputs (compressed files) of four lossless compressions algorithms: JPEG-LS and JPEG-2000 algorithms are image-dedicated algorithms, while 7z and Bzip2 algorithms are general-purpose algorithms. The relationship between the result of randomness tests and the compression ratio was investigated. This paper reports the important relationship between the statistical information behind these tests and the compression ratio. It shows that, this statistical information almost the same at least, for the four lossless algorithms under test. This information shows that 50 % of the compressed data are grouping of runs, 50% of it has positive signs when comparing adjacent values, 66% of the files containing turning points, and using Cox-Stuart test, 25% of the file give positive signs, which reflects the similarity aspects of compressed data. When it comes to the relationship between the compression ratio and these statistical information, the paper shows also, that, the greater values of these statistical numbers, the greater compression ratio we get.

Corresponding Author:

Kamal Ahmed Mulhi Al-Khayyat,
Department of Computer Science,
Kulliyah of Information and Communication Technology,
International Islamic University Malaysia, 53100 Jalan Gombak, Kuala Lumpur, Malaysia
Email: kamal_amk@yahho.com

1. INTRODUCTION

Compression is a method of reducing the size of a file for a purpose, such as saving a space, utilizing bandwidths, or increasing transmission speeds.

Compression algorithms works by removing the redundancy in data, which is why further compressing compressed data seems impossible [1]. Different approaches deal with this redundancy differently, for example, encoding the similar sequence of data (runs) in a shorter form, encoding the differences between the adjacent neighbors, and use a dictionary so that the similar sequence of data will be substituted by an index in that dictionary.

The compressed data, therefore, contains almost no redundant data, and is regarded to be random. The ordinarily meaning of randomness in statistical literature [2] or in the context of generation of random number [3] refers to the numbers of independent and identically distributed (i.i.d), or uniformly distributed numbers.

The authors in [4] used the NIST and Diehard tests to confirm the randomness of the compressed data. Their aim is to check the quality of the compressed data as a random number generator. The researchers tested the randomness of five corpora, and each was then tested as a binary sequence. In our work, we prefer converting the compressed data into unsigned integer of one-byte length, as this format is suitable for the software we selected for use. Any other format that reflects the positions or the magnitude of the data is still

accepted to serve as a measure of randomness [2]. For the tests, four non-parametric randomness tests were used.

This paper investigates the various relationships between the compression ratio and the selected randomness test, and between the compression ratio and statistical information accompanied by these tests. The paper, in Section II, briefly describe the randomness tests that were used, the sample images, and the compression algorithms applied to obtain the compressed data. In Section 3 (empirical study), we listed the relationships between the compression ratios and randomness tests and their corresponding statistical information. Section 4 summarizes the findings of the experiments

2. METHODOLOGY

Four randomness tests were used to test the randomness of the compressed data files. These data files were generated by applying four lossless algorithms on 49 gray scale images. The following subsections will describe the sample data, compression algorithms, and the randomness tests.

2.1. Compressed Data Files for Tests

Total of 49 images with a standard size of 512x512 in a raw format (PGM) were used. These gray images contain landscapes, satellite, human, animals, vehicles, buildings, and artificial images to allow the compressor algorithms to yield very different compression ratios, Figure 1 shown the sample images, which were assigned numbers 1-49. The variations in compression ratios are important to mitigate bias in testing and the accuracy of the relationships.

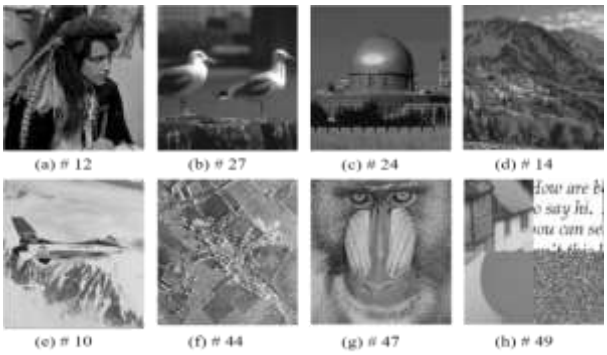

Figure 1. Samples of images used

2.2. Randomness Tests

Four randomness tests were used. These tests are commonly used to determine the randomness of the time series data [3].

The tests were used in a twofold manner; the p-values of the tests were used as a measurement of the randomness, while the statistical information–which is used to calculate the p-values –were used as well. This statistical information helps us understand the structure of the compressed data files in the context of the four tests, and their corresponding relationship(s) to the compressed data.

2.2.1. Runs test (Wald–wolfowitz Runs Test)

This test is named after Abraham Wald and Jacob Wolfowitz [5]. If we have n observations, the median value will be used as a threshold to identify two types of outcomes; the outcomes above and below the median, denoted as $n_1$and n_2, therefore, $n = n_1 + n_2$.

The expected runs for random data is computed as follows:

$$E(R) = \frac{2n_1n_2}{n} + 1 \quad (1)\backslash$$

And the variance:

$$V(R) = \frac{2n_1n_2(2n_1n_2-n)}{n^2(n-1)} \quad (2)$$

Then, the statistics test:

$$Z_R = \frac{(R - E(R))}{\sqrt{V(R)}} \quad (3)$$

R in the above equations is the total runs of both the runs of the above sequences and the runs of the below sequences. The significance level α used for all tests was 0.05, so if $Z_R > 1.96$, we reject the null hypotheses, and the data files seems to be non-random [6].

Besides using the result of the test in terms of p-value, we also used the total runs, defined as the statistical information, and when divided by the size of the compressed data, the result is defined as the statistical ratio, or ratio for short. A ratio in this case is the percentage of the number of runs found in the compressed data comparing it to its size. This naming convention will be adhered to for the rest of the randomness tests.

2.2.2. The Difference-sign Test

This test counts the number of values of i such that $y_i > y_{i-1}, i = 2, \dots, n$, or the number of times the difference series $y_i - y_{i-1}$ is positive and assign it to S, while n is the number of observations.

The mean and the variance for the random data are:

$$\mu_s = E(S) = \frac{1}{2}(n - 1) \quad (4)$$

$$\sigma_s^2 = Var(S) = (n + 1)/12 \quad (5)$$

The null hypotheses are rejected when:

$$\frac{|S - \mu_S|}{\sigma_S} > \Phi_{1-\frac{\alpha}{2}} \quad (6)$$

Where $\Phi_{1-\alpha/2} = 1.96$, when $\alpha = 0.05$, [7]. Besides the indicator of randomness represented by p-value, the number of positive signs (S) was used as a statistical information along with its ratio.

2.2.3. The Turning Point Test

If $y_1, \dots, y_n$ is a sequence of observations, there is a turning point at time $i, 1 < i < n$, if $y_{i-1} < y_i$ and $y_i > y_{i+1}$ or $y_{i-1} > y_i$ and $y_i < y_{i+1}$.

As T represents the count of turning points, the mean and the variance for a random data are as follows:

$$\mu_T = E(T) = 2(n - 2)/3 \quad (7)$$

$$\sigma_T^2 = Var(T) = (16n - 29)/90 \quad (8)$$

The null hypotheses are rejected when:

$$|T - \mu_T|/\sigma_T > \Phi_{1-\alpha/2} \quad (9)$$

Where $\Phi_{1-\alpha/2} = 1.96$ when $\alpha = 0.05$ [7]. This test was published by Irénée-Jules Bienaymé in 1874 [8]. The counts of both turning points T as a statistical information and its corresponding ratio were used.

2.2.4. Cox-stuart Test

Cox-Stuart tests is used to discover a positive or negative trend in data. The data are grouped into pairs $(X_1, X_{1+c}), (X_2, X_{2+c}), \dots, (X_{n-c}, X_n)$, where $c = n/2$ if n is even, and $c = (n + 1)/2$ if n is odd.
The signs are defined as follows:

$$sign(X_i, X_{i+c}) = \begin{cases} + & if\ X_i < X_{i+c} \\ 0 & if\ X_i = X_{i+c} \\ - & if\ X_i > X_{i+c} \end{cases}$$

Let T be the sum of positive signs, since $T \sim Bi(n, 0.5)$, $p - value$ will be a cumulative probability function for binomial distribution, and the null hypotheses will be rejected if the $p - value$ is less than α. More details can be found in [9]-[11]. The test is published in 1955 by D. R. Cox and; A. Stuart [12]. We used, besides the p-value, both the count of positive comparisons as a statistical information and its ratio.

2.3. Compression Algorithms

We used four lossless compression algorithms; two are dedicated to images, and two are general purpose algorithms:

- JPEG-LS is lossless image compression. It works by predicting the next pixel value basing on the MED (Median Edge Detection) technique, and uses Glomb-Rice for encoding [13]-[15]. The implementation provided by the Columbia University written in C language was used in this work.
- JPEG-2000 is a wavelet-based lossy-to-lossless transform coder [16]. IrfanView can be used to apply JPEG-2000.
- 7z uses LZMA (Lempel–Ziv–Markov chain algorithm) algorithms, which includes an improved LZ77 and range encoder. 7-ZIP software is used to implement this algorithm.
- Bzip2 (Bz2 for short) concatenates RLE, Burrows-Wheeler transform, and Huffman coding. It also uses the 7-ZIP software.

3. EMPIRICAL STUDY

All the tests were conducted in the R code version 3.3.1. The open source package "randtests", which includes all the tests, was used in our work. The 49 grayscale images were compressed by the compression algorithms using the software mentioned earlier. The average of compression ratios is given in Tabel 1.

In this small and varied sample data, 7z and Bzip2 were better in compression ratios, despite 7z and Bzip2 being general purpose algorithms, while JPEG-LS and JPEG2000 are image compressors. For a large data set, this may differ [17], [18].

Table 1. Average of Compression Ratios

Compression Algorithm	Compression Ratio
JPEG-LS	1.83
JPEG2000	1.64
7z	2.06
Bzip2	2.00

3.1. The Randomness Tests Results

Each randomness tests mentioned in subsection 2.2, were applied on the outputs of the compressors (subsection 3.3). The number of files that passed the randomness tests (their p-values >0.05) are shown in Fig. 2. For the run test, files compressed by Bzip2 (Bz2) shows only 3 files out of 49 that passed the test, and 19 files passing the Cox-Stuart test. In the case of the turning point test, JPEG2000 has lesser number of files that passed the test, whereas most of the files passed the Cox-Stuart test.

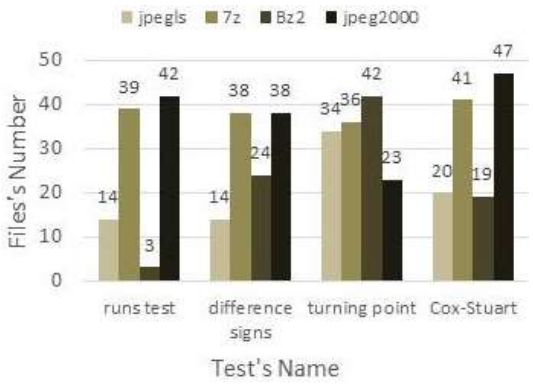

Figure 2. Number of files that passed the tests

The compressed files using JPEG-LS show similar results in runs test, difference signs, and Cox-Stuart test. In terms of the turning points, JPEG-LS comes after JPEG-2000 and near 7z in lesser number of files that passed the test.

3.2. Randomness and Compression Ratio

We did not find any significant relationship between the compression ratio and the results of the randomness tests. Table 2 shown the correlation results.

Table 2. Correlation between Randomness and Compresseion Ratio

Ratios	Compression Algorithms			
	jpegls	*7z*	*Bz2*	*Jpg2k*
RUNS	-0.319	0.015	-0.190	0.317
DIFFERENT SIGNS	-0.224	0.160	-0.303	0.098
TURNING POINT	0.057	-0.032	-0.199	0.287
COX-STUART	0.029	-0.116	0.059	-0.207

3.3. Utilizing Statistical information

It is obvious from Table 2, that there are no significant relationships between the randomness and the compression ratio. We pointed out earlier (subsection 2.2) that statistical information generated by the tests will be used to understand the structures of compressed data and their relationships to the compression ratio.

This information includes:

- Number of runs (NUM. OF RUNS), this number is used in runs test as (R) in equation (3).
- Number of positive signs (NUM. OF P_SIGNS_D_S) resulting from comparing the adjacent numbers, this number is used in deference-sign test as (S) in equation (6).
- Number of turning points (NUM. OF T_POINTS), which are used in turning points test as (T) in equation (9).
- Number of positive signs (NUM. OF P_SIGNS_COX), resulting when the second half pair wised with the first half. The symbol (T) was used.

This statistical information, when divided by their compressed file sizes, results in their ratios (or statistical ratios).

It should be pointed out that all these ratios are nearly equal (after rounding to 2 decimal points), regardless of whether their accompanied files passed or fail the randomness test. These values are:

- Ratio of number of runs (# RUNS): 0.50 (or 50% of the file size)
- Ratio number of positive signs in difference-sign test (# PS_DS):0.50
- Ratio number of turning points (# TURNIGN): 0.66
- Ratio number of positive signs in Cox-Stuart test (# PS_COX):0.25

3.3.1. Correlations between Ratios and Randomness

The correlations between the ratios and the randomness is shown in Table 3. The correlations now differ than of Table 2, no correlations shown in Table 2, significant correlations are while there were shown in Table 3.

Table 3. Correlations between the Ratios and Randomness

Ratios	Compression Algorithms			
	jpegls	*7z*	*Bz2*	*Jpg2k*
# RUNS	-0.293	-0.373	0.387	-0.779
# PS_DS	-0.630	0.312	0.328	-0.839
# TURNING	0.103	0.464	0.219	-0.846
# PS_COX	-0.353	-0.223	-0.109	-0.210

JPEG-2000 shows strong negative relationships between the ratios of number of runs, number of positive signs (difference-sign test), and the number of turning points with the corresponding randomness tests.

JPEG-LS shows the moderate negative relationship between the ratio of turning points and randomness, whereas 7z files shows the moderate positive relationships for the same test.

3.3.2. Correlations between Ratios and Compression Ratio

When taking the correlations between ratios and the compression ratio, JPEG-LS files are the only type that shows strong positive relationships between ratios of number of runs, number of positive signs (difference-sign test), and number of turning points, with the compression ratio, which are completely logical relationships. Bzip2 files shows negative strong relationship only between the ratio of positive signs (difference-sign test) and the compression ratio. Table 4 shown these relations.

Table 4. Correlations between Ratios and Compression Ratio

Ratios	Compression Algorithms			
	pegls	*7z*	*Bz2*	*Jpg2k*
# RUNS	0.663	-0.055	0.379	-0.201
# PS_DS	0.704	-0.225	-0.816	-0.053
# TURNING	0.680	-0.156	-0.503	-0.204
# PS_COX	-0.126	0.053	-0.231	0.003

3.3.3. Statistical Information and Compression Ratio

When it comes to the correlations between statistical information and the compression ratio, it is clear that a very strong negative relationship is held inside each compressed file type, without exceptions.

This relation is straightforward, i.e., the more number of statistical numbers, the less compression ratio we obtain. JPEG-2000 shows the nearly perfect relationships than 7z. Table 5 shown these relations.

Table 5. Correlations between Statistical Information and Compression Ratio

Ratios	Compression Algorithms			
	jpegls	*7z*	*Bz2*	*Jpg2k*
NUM. OF RUNS	-0.862	-0.907	-0.891	-0.981
NUM. OF P_SIGNS_D_S	-0.877	-0.907	-0.898	-0.981
NUM. OF T_POINTS	-0.862	-0.907	-0.896	-0.981
NUM. OF P_SIGNS_COX	-0.881	-0.907	-0.898	-0.981

4. CONCLUSION

There are no direct relationships between the results of randomness tests and the compression ratio, but the experiments show a strong relationship between the statistical ratios and the compression ratio for only the JPEG-LS files, except for the Cox-Stuart test.

A nearly perfect relationship was detected between the compression ratio and the ratios inside each compressed data generated by the same algorithm, which means that given two images of the same size, if compressed by the same algorithm, the compression ratios of both will conform to the statistical information of any of the four tests.

JPEG-2000 files show a strongly negative relationship between the statistical ratios and randomness results, except for the Cox-Stuart test, which means that higher ratio values result in lower p-values.

The results proved that the statistical information and statistical ratio are more beneficial when compared to the compression ratio. The results assured that, representing random files using integer representation shows that, the file harder to deal with, because of the high frequency within its data.

REFERENCES

[1] D. Salomon and G. Motta, Handbook of data compression, Fifth Edit. Springer, 2010.
[2] Y. Wang, "Nonparametric Tests for Randomness," no. May, pp. 1–11, 2003.
[3] J. S. Rustagi, "Some Tests of Randomness with Applications," DTIC Document, 1981.
[4] W. Chang, B. Fang, X. Yun, S. Wang, and X. Yu, "Randomness testing of compressed data," arXiv Prepr. arXiv1001.3485, 2010.
[5] A. Wald and J. Wolfowitz, "On a test whether two samples are from the same population," Ann. Math. Stat., vol. 11, no. 2, pp. 147–162, 1940.
[6] Panik MJ (2012) Statistical Inference: A Short Course, John Wiley & Sons, Hoboken, NJ,2012.
[7] P. J. Brockwell and R. A. Davis, Introduction to Time Series and Forecasting , Second Edition Springer Texts in Statistics.
[8] I.-J. Bienaymé, "Sur une question de probabilités," Bull. Math. Soc. Fr, vol. 2, pp. 153–154, 1874.
[9] R. H. McCuen, Modeling Hydrologic Change: statistical methods, vol. 104, no. 10. 1995.

[10] E. Lehtinen and K. Pom, "Statistical Trend Analysis Methods for Temporal Phenomena SKI Report 97 : 10 Statistical Trend Analysis Methods for Temporal Phenomena," no. April, 1997.
[11] C. Y. Man, "Applications of Non-Parametric Statistics," no. September, pp. 29–38, 1987.
[12] D. R. Cox and A. Stuart, "Some quick sign tests for trend in location and dispersion," Biometrika, vol. 42, no. 1/2, pp. 80–95, 1955.
[13] M. J. Weinberger, G. Seroussi, and G. Sapiro, *"LOCO-I: A low complexity, context-based, lossless image compression algorithm,"* in Data Compression Conference, 1996. DCC'96. Proceedings, 1996, pp. 140–149.
[14] M. J. Weinberger, G. Seroussi, and G. Sapiro, *"From logo-i to the jpeg-ls standard," in Image Processing*, 1999. ICIP 99. Proceedings. 1999 International Conference on, 1999, vol. 4, pp. 68–72.
[15] M. J. Weinberger, G. Seroussi, and G. Sapiro, "The LOCO-I lossless image compression algorithm: principles and standardization into JPEG-LS," *IEEE Trans. Image Process.*, vol. 9, no. 8, pp. 1309–1324, 2000.
[16] D. Tabuman and M. Marcellin, "JPEG2000: Image Compression Fundamentals, Standards and Parctice." Norwell, MA: Kluwer, 2002.
[17] H. Stögner, G. Weinhandel, and A. Uhl, "Experiments on Improving Lossless Compression of Biometric Iris Sample Data," Syst. Signals Image Process. (IWSSIP), 18th Int. Conf. 2011, 2011.
[18] G. Weinhandel, H. Stogner, and A. Uhl, *"Experimental study on lossless compression of biometric sample data,"* in Image and Signal Processing and Analysis, 2009. ISPA 2009. Proceedings of 6th International Symposium on, 2009, pp. 517–522.

BIOGRAPHIES OF AUTHORS

Kamal AL-Khayyat is a Ph. D candidate at IIUM (International Islamic University Malaysia). He obtaind his BSc in computer science) from the university of scincein and technology, Yemen,2000, MSc in computer science from Dr. Baba Saheb, India, 2005.He has taought various subjects of computer sciences with more than seven years of experience as a lecturer. His main interests are, theory of computation, image processinga and data compression

Imad Al-shaikhli is a professor and the head of research at IIUM (International Islamic University Malaysia). He is also a lecturer at the Faculty of Information and communication Technology. He is a IEEE senior member, obtained his BSc (Hon) in Mathematics, MSc in Computer Science from Iraq, and Ph. D degree from Pune University, India, 2000. He has been the editor in chief of International journal on Advanced Computer Science and Technology Research since 2011 now, and the general chair of the international conference on Advanced Computer Science Applications and Technologies since 2012 till now. He obtained a US patent for his work with his Ph. D student on smart traffic light with accident detection system on 2nd December 2014. Prof. Imad has published more than 100 papers, journals and book chapters in addition to three books.

Prof. Dr. Vijayakumar Varadarajan. Currently, he is a Professor of school of computer science and Engineering at VIT University, Chennai, India. He has more than 16 years of experience including industrial and academic. His research interests span in computational areas covering grid computing, cloud computing, computer networks and big data. He has completed BE, CSE and MBA HRD with First Class. He has also completed ME, CSE and MBA HRD with First Class. He completed his PhD from Anna University in 2012. He ia a reviwer in IEEE Transactions, Inderscience and Springer Journals. He has initiated a number of international research collaboration with university in Eurpoe, Australia, Africa and North America.