Restatement of Information Statistics Through the Concept of Unit Histogram

Radu D. RUGESCU
"Politehnica" University of Bucharest
313 Splaiul Independentei, sector 6, Bucuresti
rugescu@yahoo.com

***Abstract*—The restatement of classical statistical analysis is proposed with the new unit histogram concept that increases with one order of magnitude the amount of statistical information extracted from experimental data. The new statistical technology is particularly suited for scarce sample populations, when it increases with one order of magnitude the extracted information. Hidden features would remain inaccessible otherwise. The method could prove useful also for signal processing, when a high level of accuracy in details rendition is required. The efficiency of the method is demonstrated by the experimental measurement of the combustion heat of rocket propellants. Populations as small as of six elements are thus successfully processed and statistically characterized, while the regular statistics could only be applied to much larger populations.**

***Index Terms*—statistical analysis, histogram, travelling bin histogram, combustion heat of rocket propellants**

The restatement of classical statistical analysis is proposed with the new unit histogram concept that increases with one order of magnitude the amount of statistical information extracted from experimental data. The new statistical technology is particularly suited for scarce sample populations, when it increases with one order of magnitude the extracted information. Hidden features would remain inaccessible otherwise. The method could prove useful also for signal processing, when a high level of accuracy in details rendition is required. The efficiency of the method is demonstrated by the experimental measurement of the combustion heat of rocket propellants. Populations as small as of six elements are thus successfully processed and statistically characterized, while the regular statistics could only be applied to much larger populations.

I. INTRODUCTION

Histograms were introduced for long to provide a means of estimating the character of a statistical distribution of data (population), especially in connection with the Gaussian or normal distribution (Matusita, 1955). They are the regular histograms with fixed bins, almost exclusively used in different types of data and signal processing (Cook *et al.*, 1963), (Amabile, 1989), (Plackett, 1965), (Austin *et al.*, 1991), (Obraztsov *et al.*, 2006). We address here a common set of such data from measurements, or items, composing a univariate population. For bivariate data the same regular technique is used (Hamalainen, 2004) and for multivariate problems the main concern (Rubner *et al.*, 1998) is in the connectivity of individual columns, especially when their variance is different. Ingenious techniques were imagined (Serratosa *et al.*, 2005), all of them based on the fixed bin sampling method (Fuchs *et al.*, 2007). For example, the compressed histograms for selectivity estimation in (Fuchs *et al.*, 2007) are also of regular type. These applications are usually oriented towards large sample populations, where data grouping does not discountenance the analysis (Lu *et al.*, 2005), (Bairamovic *et al.*, 2005), (Gibbons *et al*., 2002) and, up to a point at least, seems unavoidable for a correct statistical image of the population. The authors in (Lam *et al.*, 2005) apply their development to fixed bins also and mobile windows are used only in updating the population.

Specific filtering techniques are often used (Austin *et al.*, 1991) to extract the main frequencies of variable signals (Lobos *et al.*, 1996), (Berghaus *et al.*, 2001). The best known methods are based on digital filters of Hamming type (Hamming, 1962), (Jackson *et al.*, 1986), Hann or Blackman (Lobos *et al.*, 1996). Very documented presentations of these techniques are found, among other sources, in (Berghaus *et al.*, 2001), (Ciochina, 2006), (Allen *et al.*, 1977), (Harris *et al.*, 1978), (Griffin *et al.*, 1984), (Brunelli *et al.*, 1995).

These methods are based in totality on the superposition of modifying windows, based on a synthesised transfer function with a given weighing law that exaggerate the frequencies around an arbitrary basic or ***desired*** frequency ω, in order to underline any existing main modes, and diminish the other parts. The start point in developing the filtering techniques is the requirement to reduce the noise.

Even when dealing with large populations and with vectors of data (Juhee, 2006), (Gish *et al.*, 1985), from image or sound (Gish *et al.*, 1985) processing, like in person identification algorithms for example (Brunelli *et al.*, 1995), the construction of the histograms is performed on the same classical rules of data grouping. Filters, including digital, are imagined for real time application. Each moment the signal is then intentionally distorted, within a band of frequencies.

The Hamming and Hann windows, for example (Basseville, 1989), (Hamming, 1962), (Ciochina, 2006), multiply the harmonics around the desired frequency ω by the set of coefficients, depending on the parameter α,

$$w_H(n) = \begin{cases} a - (1-a)\cos\dfrac{2p}{N}(n+\dfrac{1}{2}), & n \in [\overline{1, N-1}] \\ 0, & \textit{in the rest}. \end{cases}$$

The number of covered harmonics N is the length of the filtering window, only related to frequencies from the signal. The slight difference between the two filtering windows is that the Hamming corresponds to α=0.54 while the Hann is for α=0.5 and nothing more (Ciochina, 2006). Despite this minor difference, them were attributed different names.

The new method we present here induces in fact true reversed effects, each value of the statistical set being reproduced independently and in its real size, without any alteration, amplification or adjustment, opposite to the case of usual digital filters.

Perhaps it represents mainly a ***true-filtering*** technique, by which the influence of other neighbouring values is dismissed and the true personality of each encountered measurement (reading) is immediately considered in the histogram, without waiting to be grouped with other items. Once the very detailed histogram is accomplished, it remains to be further analysed and the effect of noise to be segregated by extra means, depending on the resulted, real statistical distribution.

II. Regular histograms

The statistical information is processed by grouping the data of the set into a number m of equally sized (width $2b$) collecting bins and counting the *appearance* h_i of any individual reading v_i within every j bin, appearance which could equal either 0 or 1,

$$h_i(v_i;\ c_j, b) = \mathrm{int}\frac{v_i}{(c_j - b)} - \mathrm{int}\frac{v_i}{(c_j + b)}\ . \tag{1}$$

The symbol *int* denotes integer truncation. Besides its given width $2b$, each j bin is defined by its centreline position c_j, starting with the first centreline c_1, also given,

$$c_j = c_1 + 2b\,(j-1) \tag{2}$$

and the hierarchy $v_n + a < 2\,(v_1 + a - b)$, based on a constant a, conveniently chosen, is assumed.

Referring to scalar sets for example, which are the basic stuff of usual experimental measurements, after summing the individual histograms (1) over the entire set

$$h(c_0, b) = \sum_{i=1}^{n}\sum_{j=1}^{m} h_i(v_i\,; c_j, b)\,, \tag{3}$$

a bar diagram is resulting, or the *histogram* (Fig. 2) of the set. Its aspect hangs on two empirically chosen constants: the width of the bins $2b$ and the location of the first bin, namely its centreline, c_1 (Rugescu, 2007b).

It is concluded that histograms play a central role in experimental data evaluation for errors and confidence, in medical care statistics and equally in image and sound transmission, extensively used today in deep space communications and in person, signature, fingerprint, or voice recognition (Cook *et al.*, 1963), (Amabile, 1989), (Coolidge, 2006), (Juhee, 2006), (Brunelli, 1995), (Gish *et al.*, 1985). To gain clear information about the distribution by classical histograms, the size n of the population must exceed 20-40 items (Cook *et al.*, 1963), (Brdiczka *et al.*, 2006), due to the grouping process involved.

Independent of the grouping process, the mean and standard deviation over all values of the set

$$\mu_0 = \sum_{i=1}^{n} v_i \Big/ n\,,\quad \sigma_0 = \sqrt{\sum_{i=1}^{n} (v_i - \mu_0)^2 \Big/ n}\,, \tag{4}$$

computed either with initial data or with weighted centrelines of the histogram, show the magnitude of errors and the confidence in the experiment (Apostolescu *et al.*, 1979). The grouping of data into fixed bins distorts however the existing information. Identical histograms could prove belonging to basically different distributions (Fig. 1), engaging this way inherent inaccuracy. For large populations, the trailing error is acceptably small, especially when some corrections are introduced (Cook *et al.*, 1963). When the data go scarce however, grouping is no more possible and the regular histograms become unachievable. Consequently the representation of a scarce data set by a histogram becomes out of practical reach. Data sequencing remains hidden.

III. Unit histograms

A considerable improvement is introduced when, instead of assuming given positions to m fixed collecting bins, a unique sliding bin of equal width $2b$ is continuously moved along the data set. Each time a new reading enters the travelling bin window, the frequency is increased with one unit and vice versa, when it exits the window. The individual histogram of every value v_i, centred on its very own reading v_i, namely the function of double unit-step type with the edges on v_i-b and v_i+b,

$$h_i(v\,; v_i, b) = \mathrm{int}[v/(v_i - b)] - \mathrm{int}[v/(v_i + b)]\,\big|\,v \in \Re,\ i = 1, n \tag{5}$$

will be added, up to the current position v of the sliding bin,

$$h(v; b) = \sum_{j=1}^{i} \mathrm{int}[v/(v_j - b)] - \mathrm{int}[v/(v_j + b)] \tag{6}$$

to give finally the total *centreline wake* (CLW) histogram $h(b)$.In these formulae *int* denotes truncation again and the hierarchy $v_n + b < 2\,(v_1 - b)$ is evoked.

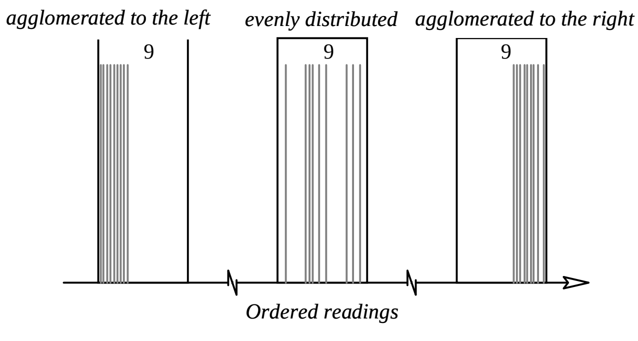

Figure 1. Insensitivity of fixed bins to internal distribution**..**

Example of insensitivity for nine items grouped in a bin.

When not directly met, the set $\{v_i \,|\, i = 1, n\}$ will be translated with a constant $a > v_n - 2v_1 + 3b$ (Rugescu, 2007b) to comply with the required hierarchy, fact always possible. The alternative of raising hi right at the left margin of the travelling bin (not at centreline) for a new vi reading encounter at its right side (sliding bin pattern SBP) (Rugescu, 2007a) proves less useful. The new histogram is sensible to any change in an individual position of a reading, eliminating completely the intrinsic inaccuracy of the regular histograms. Furthermore, when the mean and standard deviation are short-computed by replacing the original set with the centreline of the fixed bins, the intrinsic inaccuracies transmit into them. This doesn't happen with the new method and the standard deviation is computed more accurately. The following numerical examples show that the amount of statistical information revealed through the CLW histogram doubles the original scarce data information, besides eliminating the intrinsic inaccuracy. Thus the travelling bin histogram produces no alteration of the initial data and secures a clearly enriched and personalised statistical picture.

IV. EXPERIMENTAL EXAMPLE AND RESULTS

The new statistics was first used to analyse scarce laboratory measurements of the isochoric heat of combustion Qv of a solid homogeneous rocket propellant (Rugescu, 2007b), performed in a common 282 cm3 calorimetric bomb under vacuum (Rugescu, 2007a). Samples from a nominal blend of triple base propellants of genuine origin were carefully investigated. Due to technical difficulties and other causes the population vi was relatively scarce. To induce enough confidence in the measurements a strong statistical analysis was required that faced the problem with the very limited amount of data: 6 readings in the first run and 17 readings in the second. Under these circumstances regular histograms are of no use and a different approach is required. From there the method of the travelling bin had emerged, proving however a reliable tool for much wider applications.

TABLE 1. SAMPLE MEASUREMENTS IN THE SECOND EXPERIMENT ON THE HEAT OF COMBUSTION

Measurement no. i	Reading v_i (*cal/g*)
1	853.62
2	855.10
3	855.74
4	855.95
5	856.37
6	857.85
7	858.70
8	859.33
9	859.44
10	859.55
11	859.76
12	859.97
13	860.60
14	860.81
15	861.03
16	862.93
17	868.01

The set of n=17 delivered values (vertical short grey lines in Fig. 2) are here used to first compare the new with the classical method. The ordered values (increasing) of the experiment are given in table 1.

This scarce set is considered as a test for the efficiency of the new statistics, which are compared, first by grouping the data in seven collecting bins, with the equal width of $2b$=2.33 units (cal/g in this case). The first bin starts at its centreline location c_1=854.36 namely at position 853.20 cal/g. As usual, the resulting histogram has a visibly coarse aspect (Fig. 2), where the intrinsic details are hidden, due to this process of grouping.

Furthermore, variation in the size and position of the fixed bins produce large variations in the aspect of the regular histogram. With a travelling bin of 2.33 units in width a new CLW histogram type is now built (Fig. 3). Hidden details in the first histogram are now impressively revealed.

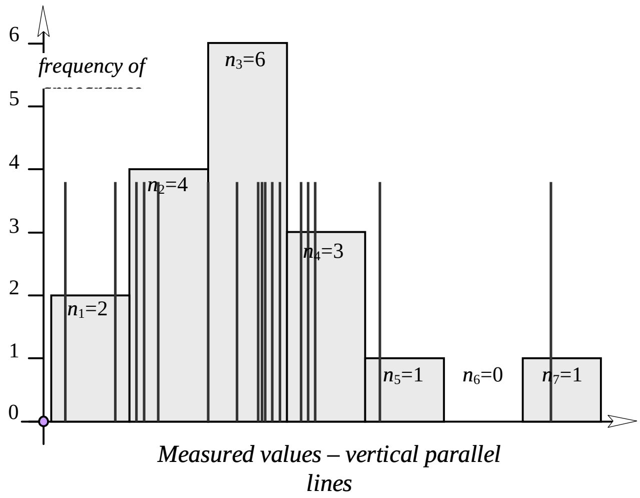

Figure 2. Regular histogram for the n=17 sample.

Among these details, it is well revealed in Fig. 3 for example the layout of a bimodal statistical distribution of measurements, namely a melange of two separate normal distributions that may be analysed independently.

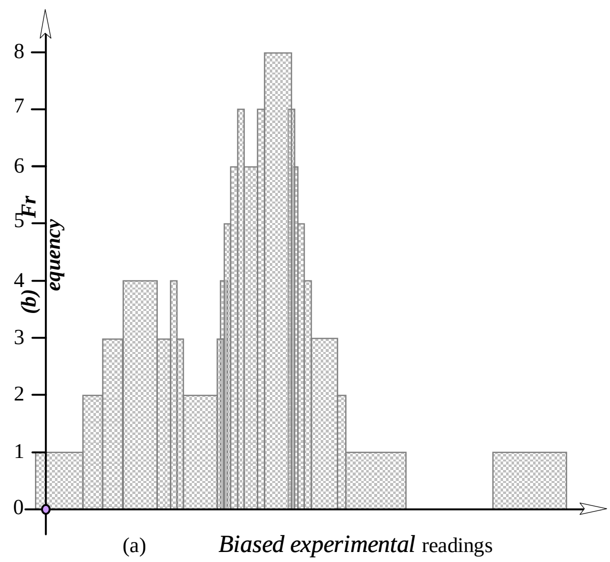

Figure 3. Multiple unit CLW histogram for data set in Table 1.

To perform this separate weighing for the bimodal analysis, the set is split into the left six readings and the right twelve ones. Thus the two Gaussian distributions are revealed and represented into the same drawing (Fig. 4).

Considering the right part of the set only of twelve items, the isolated reading from its right margin seems to be clearly affected by noise and two distributions were consequently computed and drawn. The first, comprising 12 elements, includes the extreme right element, while the other, with 11 elements only, excludes the right hand, heretic reading. This isolated value proves to produce a minor shift of 0.668 units of the mean of the partial set, while the modification of the standard deviation looks very large, almost double in the presence of the irregular reading. The dismissing of the isolated value remains rather arbitrary however.

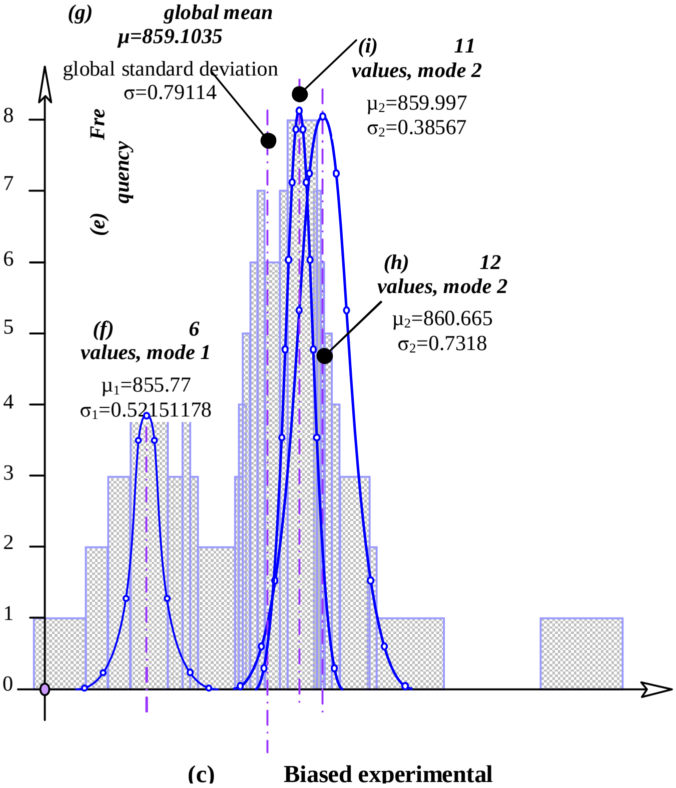

Figure 4. New CLW histograms reveal the bimodal behaviour.

Returning to the general aspect of the new histogram, comparison of the diagrams in Fig. 2 and 3 gives a useful evaluation of the improvement introduced. It was eventually observed that the accuracy of thermochemical computer simulations was highly improved when these values were used to determine the standard enthalpy of formation of the cellulose ester with various degrees of nitration, largely used as main component in the colloidal propellants.

V. LIMIT EXAMPLE WITH SIX READINGS

A second example shows further how a very good CLW histogram can be built for a set of only 6 readings, where the standard histogram can not this time be built at all. A different measurement action upon the heat of combustion at constant volume was thus performed and brings into attention the readings in Table 2, containing only 6 elements.

A usual histogram can not be conveniently attached to this very small amount of data. To force however a regular type representation, three fixed collecting bins with the width 2σ are still selected with the result in Fig. 5. The window width was adopted in this drawing for the sake of uniformity with the previous examples.

TABLE 2. MEASURED VALUES IN THE SECOND EXPERIMENT ON THE HEAT OF COMBUSTION OF ROCKET PROPELLANTS.

Sample no.	N_2, %	Q_v, cal/g
1	12,04	987,5
2	12,03	987,8
3	11,95	983,6
4	11,96	970,0
5	11,86	961,0
6	11,89	971,2

Neither the known condition to have at least six collecting bins in the histogram (Cook et al., 1963) could be met with these only six values. It is clear to say that no information could thus be driven from the regular histogram with these six readings. No extra measurements were available under the same experimental conditions and the only way to investigate the distribution is by a new method.

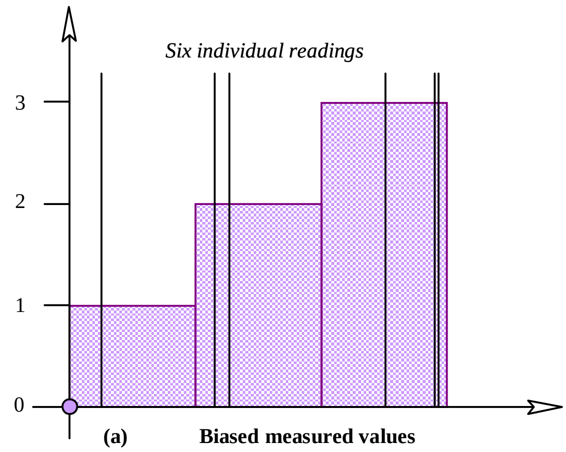

Figure 5. Formal, non-relevant standard histogram of the n=6 data set.

Let apply now to this set the new formula of representation, which lets each individual value to play its own role within a $2b$ width statistical correlation. This manner the new histogram is built again (Fig. 6).

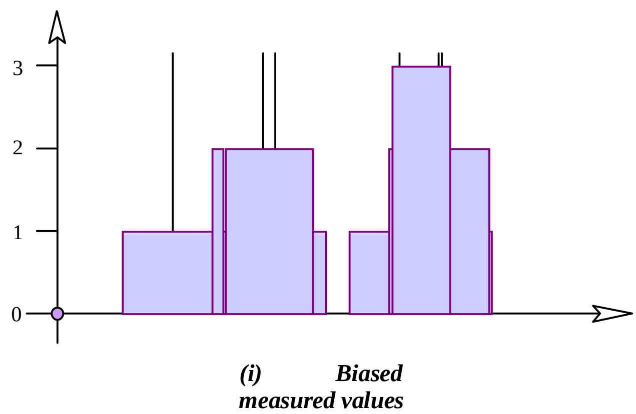

Figure 6. Detailed CLW histogram of the n=6 data set (unit histograms with 2σ width).

This is the new CLW type histogram that offers instead an unexpectedly detailed information about this very small distribution. It reveals for example that a dual mode distribution is again present, which suggests that a repeated error in the measurements had been embedded. The left and right parts of the histogram resemble two adjacent normal distributions superposed. Considering each of them in the overall distribution up to an acceptable degree as a Gaussian normal distribution, two distinct sets are considered within the diagram. This example could say a lot about the extracting power of the CLW histogram. It proves especially suited as a tool for very scarce statistical populations. Each individual reading produces two modifications in the CLW histogram, one at its entrance and the second on its exit from the travelling bin window. As far as a group within a bin of the standard histogram involves at least five readings, the CLW histogram is an order of magnitude more detailed.

Data in Fig. 6 seem to belong to two distinct sets, where the conclusion regarding the dual mode population comes from when the sliding bin has the width of 2σ for the whole distribution. The effect of smaller widths is shown below.

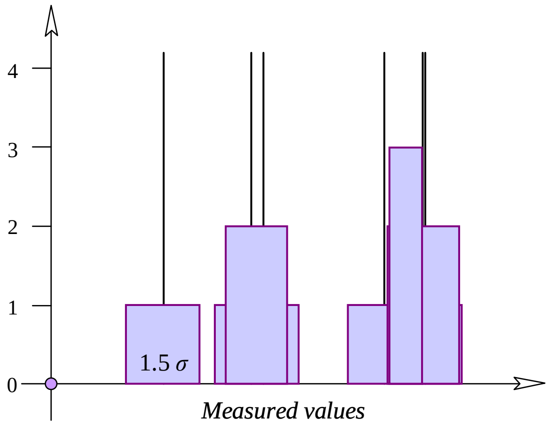

Figure 7. Same data CLW histogram (Table 2) with a 1.5σ width travelling bin.

Figure 7 shows that for a width of 1.5σ the data are splitting into three groups, meaning a lower correlation. The same preserves for a width of the travelling bin of 1σ (figure 8).

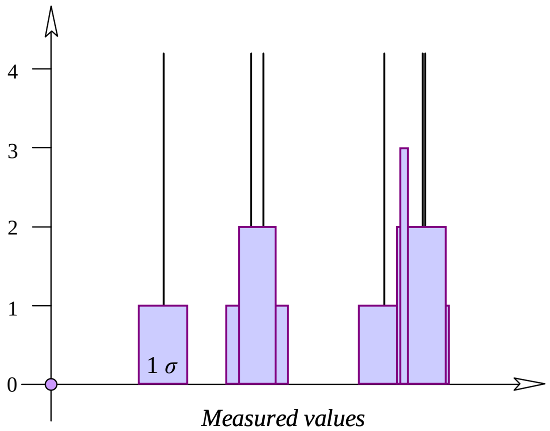

Figure 8. Same data CLW histogram with the 1σ width travelling bin. Data become looking non-correlated.

In this limit case of data scarcity the impact of the travelling bin width is important, the image of the CLW histogram is visibly changing. Consequently the question on its optimal size remains to be solved, as other aspects do, as for example the correlation properties of the CLW histograms.

VI. CONCLUSIONS

Through the performed laboratory measurements and by applying the enhanced statistical tool of the *travelling bin histogram*, the value of 859 cal/g for the heat of combustion of the A-100 triple base propellant was obtained. A good standard deviation of σ_0=0,791 cal/g appeared. A grouped histogram should not have been structured so far. Not only the new histogram was clearly structured, but it also revealed a slightly bi-modal distribution, which suggests that a small systematic error was introduced into the measurements.

The source could not have been established. Including this uncertainty, the accuracy of the measurements entered well within the requested technical accuracy. The new type of travelling bin histograms is behaving as an especially relevant statistical tool, opening a general applicability in the technology of measurements. The resulting histogram doubles the amount of information of the primary data set and raises with one order of magnitude the information of classical, regular histograms, transforming this tool into a recommendable means for the statistical analysis in general.

This method was successfully used in processing statistical sets as scarce as of six readings only, producing a clear CLW histogram, where the classical statistic is impossible. As the CLW histogram is a one variable problem, the only thing that follows is to properly choose the width of the collecting bin as a problem of optimum value. This aspect is under current development by the author, based on the so-called *distance between histograms* (Wilcox, 2003), (Brdiczka et al., 2006), (Juhee, 2006), (Yu et al., 2006). Similar extensions are expected with applications to vector or multivariate data sets, typical for complex signal processing, where similar advantages are envisaged when convenient assumptions are set. The demonstration of the utility of this natural histogram is thus considered for univariate problems as simpler and more relevant. Results from application of the new technology are obviously welcomed.

As far as the randomness in positioning of collecting bins is removed with the new travelling bin, this method reduces the statistical analysis to a one variable problem and the correlation of data is strongly emphasised, based on the very value of the individual readings. Consequently, the new travelling bin histograms are sensible to the slightest change in readings value, rendering a really *personalised problem.*

Extensions are expected in other fields like to spectral entropy application, mainly in important medical care data processing (Xiaoli, 2006), (El-Samahy *et al.,* 2006) and to bivariate S- distributions utilising copulas (Yu *et al.,* 2006) or to multivariate S-distributions (Plackett, 1965), (Serratosa *et al.,* 2005), (Brdiczka *et al.,* 2006) and vectors of data for example, where the new method could prove largely profitable.

REFERENCES

[1] Allen, J.B. and L.R. Rabiner A unified approach to short-time Fourier analysis and synthesis, Proceedings of the IEEE, ISSN 0018-9219, 65, Issue 11, 1977, pp. 1558-1564.

[2] Amabile, T. Describing distributions, Consortium for Mathematics and Its Applications, Chedd- Angier Production Co., 59 min. Videorecording, KAMU-TV, College Station, Texas USA, 1989.

[3] Apostolescu, N. and D. Taraza Basics of experimental research in thermal machines, pp. 443-472. Ed. Didactică şi Pedagogică, Bucureşti, 1979.
[4] Austin, J.D., D.W. Scott Beyond histograms: average shifted histograms, Math. Comput. Educ., 25, 1, 1991, pp. 42-52.
[5] Bajramovic, F., Ch.Grassl and J. Denzler Efficient Combination of Histograms for Real-Time Tracking Using Mean-Shift and Trust-Region Optimization, in Kropatsch, W. Sablatnig, R., Hanbury, A. (Eds.): DAGM 2005, LNCS 3663, Springer-Verlag Berlin Heidelberg 2005, pp. 254-261.
[6] Basseville, M. Signal Processing, 1989, pp. 349-369.
[7] Berghaus, D. Numerical Methods for Experimental Mechanics, Kluwer Acad. Publ., ISBN 0-7923-7403-7Dordrecht, 312 p, 2001.
[8] Brdiczka, O., J. Maisonnasse, P. Reignier and J.L. Crowley 10th Int. Conf. KES 2006, Bournemouth, UK, Oct. 9-11, 2006, Part II, Springer ISBN 978-3-540-46537-9, p. 162.
[9] Brunelli, R. and D. Falavigna Person Identification Using Multiple Cues, 17, 10, pp. 955-966, 1995.
[10] Ciochina, S. Numerical Processing of Signals, Chapter 10: Fast Algorithms to Perform Convolution and the Discrete Fourier Transform, University Politehnica of Bucharest, www.comm.pub.ro/curs/pns/cursuri/capitolul2p1.pdf (in Romanian), 2006.
[11] Cook, N.H. and E. Rabinovicz Physical Measurement and Analysis, p.29-68. Addison-Wesley Publishing Co., Reading, Massachusetts, 1963.
[12] Coolidge, F.L. Statistics: a gentle introduction, 2nd ed., Sage Publications ISBN 1412924944, Thousand Oaks, CA, 2006.
[13] El-Samahy, E., M. Mahfouf and D.A. Linkens Artificial Intelligence in Medicine 38, pp. 257-274, 2006.
[14] Fuchs, D., Z. He and B.S. Lee Info. Sci. 177, pp. 680-702, 2007.
[15] Gibbons, Ph. B., Y. Matias and V. Poosala ACM Transactions on Database Systems, 27, 3-Sept. 2002, pp. 261–298.
[16] Gish, H., J. Makhoul and S. Roucos "Vector quantization in speech coding," Proc. IEEE, 73, 11, pp. 1551-1588, 1985.
[17] Griffin, D. and L. Jae Signal estimation from modified short-time Fourier transform, IEEE Transactions on Signal Processing, ISSN 0096-3518, 32, Issue 2(1984), pp. 236-243, 1978.
[18] Hamming, R.W. Numerical methods for scientists and engineers. International Series in Pure and Applied Mathematics, McGraw-Hill Book Co., Inc., New York-San Francisco, Calif.-Toronto-London, xvii+411 p, 1962.
[19] Hamalainen, J.S. IEE Proc.-Sci. Meas. Technol., 151, 6 Nov. 2004, pp. 492-495.
[20] Harris, F.J. On the use of windows for harmonic analysis with the discrete Fourier transform, Proc. IEEE, 66, 1, pp. 51-83, 1978.
[21] Jackson, L.B. Digital Filters and Signal Processing, Boston, MA: Kluwer A.P, pp. 85-92, 1986.
[22] Juhee, S. Bootstrapping in a high dimensional but very low sample size problem, Ph. D. Thesis, Texas A&M University, USA, 2006.
[23] Lam, E. and K. Salem Proceedings of the 9th International Database Engineering & Application Symposium - IDEAS-2005.
[24] Lobos, T. and J. Rezmer Real-Time Determination of Power Systems, Proceedings of the lEEE Instrumentation and Measurement Technology Conference, Brussels, Belgium, June 4-6, 1996, pp. 756-761.
[25] Lu, Z.-M., H. Burkhardt Electronics Letters, 18th August 2005 41, 17.
[26] Matusita, KThe Annals of Mathematical Statistics, 26, 4- Dec. 1955, pp. 631-640, 2005.
[27] Obraztsov, S.M., Yu.V. Konobeev, G.A. Birzhevoy and V.I. Rachkov 2006. J. Nucl. Mat., 359, 2006, pp.263-267.
[28] Plackett, R.L. J. Amer. Statistic. Assoc., 60, No. 310, June 1965, p. 516.
[29] Rubner, Y., C. Tomasi, and L.J. Guibas A Metric for Distributions with Applications to Image Databases, pp. 59-66, 1998.
[30] Rugescu R.D. Retrieving Data Structure from Scarce Population Sample, Private presentation, Aerospace Engineering Department of Texas A&M University, College Station USA, Feb. 29, 2007.
[31] Rugescu R.D. Revista de Chimie, [ISI-Thomson Journal ISSN 0034-7752], Ed. Chiminform Data S. A., 59, 2007, pp. 579-581.
[32] Serratosa, F. and A. Sanfeliu N. Computer Sci., 3773; 10th Iberoamerican Congress on Pattern Recognition, CIARP-2005, p. 1027.
[33] Su, Y.T. and Ju-Ya Chen IEEE Transactions on communications, 48, 8Aug. 2000, pp. 1338-1346.
[34] Wilcox, R.R. Applying contemporary statistical techniques, Academic Press ISBN 0127515410, San Diego, CA, 2003.
[35] Xiaoli, Li. Lecture Notes in Computer Science/Neural Information Processing, Springer-Verlag, Volume 4234/2006, pp. 66-73. (The 13th International Conference on Neural Information Processing-ICONIP2006, HK. 2006 Oct.3-6.)
[36] Yu, L. and E.O. Voit Computational Statistics & Data Analysis, 51, pp. 1822-1839, 2006.