Supplementary material

NOTE 1 – On the use of natural or surrogate samples to control the analytical process

The frequency distribution of an analyte value in a biological sample, which reflects the state of a biological parameter, depends on the intrinsic characteristics of an individual and the conditions in which the individual is at the time of observation (*i.e.*, sample collection). The width of this distribution (assumed to be normal) is given by:

biological variance = between-subject variance + within-subject variance

or in biostatistical terminology considering a given population's homeostatic point ($\bar{x}$):

$$CV_{biological} = CV_{group} + CV_{individual}.$$

Measuring the parameter, an additional component is added to the biological variance of the value. This depends on the distribution of the error associated with the measurement method used:

total variance = biological variance + analytical variance

and again:

$$CV_{total} = CV_{biological} + CV_{analytical}.$$

The analytical variance plays a crucial role when the same sample analysis is repeated, as for instance when the sample is used for controlling the analytical process according to the Levey and Jennings adaptation of the Shewhart quality control procedure. Conversely, biological variance is pivotal when the sample has exhausted its half-life (consumed or degraded) and needs to be restored.

Indeed, the value of a new biological sample cannot be chosen but only be expected within a range whose width depends on biological variance. If this variance is large and analytical variance changes with the concentration of the analyte (*i.e.*, constant

coefficient of variation), the new control sample with a different target value imposes to completely reset the control procedure (*i.e.,* loosing historical data since referring to different conditions).

The Belk and Sunderman model (reference 3 in bibliography), formerly proposed for external proficiency testing, elegantly solves this problem by replacing the biological sample with a surrogate whose characteristics do not affect the analytical variance (commutability). The surrogate sample is prepared by adding the analyte to a natural or artificial matrix, so its concentration (to be used as the target for the process) is determined in advance. This allows the biological variance to be eliminated (because the variance of a constant is zero), or rather replaced with a component, which we might call "metrological," that is manipulable and therefore can be made negligible for the purpose of controlling the analytical process quality.

For a surrogate control sample, we can say:

$$CV_{metrological} << CV_{biological} \text{ or } CV_{metrological} \rightarrow 0$$

Thus:

$$CV_{total} = CV_{metrological} + CV_{analytical} \rightarrow CV_{total} = CV_{metrological}$$

The Belk and Sunderman model is cost-effective, standardizable, and metrologically traceable, and represents the dominant model for control materials in modern times although it is not properly acknowledged. It allows to reproduce the analyte value either in normal or pathological conditions, so it involves the preparation of multiple samples to control the same process at different levels. It is therefore intuitive that, as the process is unique, the measured values of the control levels tend to be correlated.

NOTE 2 – Symbols and properties of matrices (a gentle introduction)

This note aims to provide sufficient knowledge to understand the formulas presented in the article. It is not exhaustive of matrix algebra, and readers are encouraged to consult appropriate texts for a deeper understanding.

A matrix is defined and denoted by the bold letter **A** (or with the upper script $\mathbf{A}^{\rightarrow}$) as an orderly array of k elements in p columns and m rows (so k = m*p):

$$\mathbf{A} = \begin{bmatrix} 4 & 3 \\ 7 & 11 \\ 3 & 5 \end{bmatrix}.$$

Here are some key types and properties of matrices:

a) Scalar: a 1 x 1 matrix, *i.e.*, a single number.

$$\mathbf{A} = [5] = 5$$

b) Vector: a 1 x p matrix (one row and p columns) or mx1 matrix (m rows and one column).

$$\mathbf{A} = [3 \quad 9]\,; \mathbf{B} = \begin{bmatrix} 5 \\ 10 \end{bmatrix}$$

c) Square matrix: an m x p matrix where m = p; this matrix identifies the elements forming the diagonal of the array (in the example, the numbers “9”, “1”, and “8”).

$$\mathbf{A} = \begin{bmatrix} 9 & 2 & 3 \\ 4 & 1 & 4 \\ 5 & 7 & 8 \end{bmatrix}$$

d) Identity matrix: a square matrix where all elements are 0 except for those on the diagonal, which are 1; denoted by “**I**”.

$$\mathbf{I} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

e) Symmetric matrix: a square matrix where off-diagonal elements are mirror-symmetrical.

$$\boldsymbol{B} = \begin{bmatrix} 3 & 5 & 9 \\ 5 & 1 & 2 \\ 9 & 2 & 4 \end{bmatrix}$$

Consider now two matrices **A** = m_A x p_A and **B** = m_B x p_B:

f) Sum and subtraction: these operations proceed element by element, so it is necessary that $m_A = m_B$ and $p_A = p_B$, *i.e.*, the matrices must have the same dimensions. If this condition is met, then **A** + **B** = **B** + **A** = **C**, as in linear algebra.

$$\mathbf{A} + \mathbf{B} = \begin{bmatrix} 2 & 5 \\ 1 & 2 \end{bmatrix} + \begin{bmatrix} 3 & 6 \\ 1 & 6 \end{bmatrix} = \begin{bmatrix} 2+3 & 5+6 \\ 1+1 & 2+6 \end{bmatrix} = \begin{bmatrix} 5 & 11 \\ 2 & 8 \end{bmatrix}$$

g) Matrix product: this operation is allowed only if the first multiplier has as many rows as the columns in the second multiplier (*i.e.* $m_A = p_B$), and it results in a matrix with m_A rows and p_B columns. Therefore, **A** x **B** ≠ **B** x **A**.

$$\mathbf{A} \times \mathbf{B} = \begin{bmatrix} 1 & 2 & 4 \\ 5 & 3 & 0 \end{bmatrix} \begin{bmatrix} 2 & 6 \\ 7 & 0 \\ 3 & 9 \end{bmatrix} = \begin{bmatrix} 1*2+2*7+4*3 & 1*6+2*0+4*9 \\ 5*2+3*7+0*3 & 5*6+3*0+0*9 \end{bmatrix} = \begin{bmatrix} 28 & 42 \\ 31 & 30 \end{bmatrix}$$

$$\mathbf{A} \times \mathbf{B} = \begin{bmatrix} 2 & 6 \\ 7 & 0 \\ 3 & 9 \end{bmatrix} \begin{bmatrix} 1 & 2 & 4 \\ 5 & 3 & 0 \end{bmatrix} = \begin{bmatrix} 2*1+6*5 & 2*2+6*3 & 2*4+6*0 \\ 7*1+0*5 & 7*2+0*3 & 7*4+0*0 \\ 3*1+9*5 & 3*2+9*3 & 3*4+9*0 \end{bmatrix}$$

$$= \begin{bmatrix} 32 & 22 & 8 \\ 7 & 14 & 28 \\ 48 & 33 & 12 \end{bmatrix}$$

h) Scalar multiplication: here, the scalar “d” is multiplied by all elements of the matrix, i.e., d x **A** = d**A**.

$$\mathrm{d} \times \mathbf{A} = 3 \begin{bmatrix} 2 & 5 \\ 1 & 2 \end{bmatrix} = \begin{bmatrix} 3*2 & 3*5 \\ 3*1 & 3*2 \end{bmatrix} = \begin{bmatrix} 6 & 15 \\ 3 & 6 \end{bmatrix}$$

i) Transpose: this operation swaps rows and columns within the matrix (maintaining the order of elements), and the transpose is distinguished by the symbol "'"; thus, if **A** = m x p, then **A'** = p x m and (**A'**)**' = A.**

$$\mathbf{A} = \begin{bmatrix} 1 & 2 & 4 \\ 5 & 3 & 0 \end{bmatrix} \rightarrow \mathbf{A}' = \begin{bmatrix} 1 & 5 \\ 2 & 3 \\ 4 & 0 \end{bmatrix}$$

j) Inversion: this is a transformation of the matrix that can be assumed to correspond to 1/n for the scalar n; the inverse is denoted with the exponent "**-1**" and it is important to note that not all matrices are invertible (there are methods to determine which ones are)

$$\mathbf{A} = \begin{bmatrix} 2 & 3 \\ 6 & 7 \end{bmatrix} \rightarrow \mathbf{A^{-1}} = \begin{bmatrix} -1.75 & 0.75 \\ 1.5 & -0.5 \end{bmatrix}.$$

NOTE 3 – Calculation of the vector of means

The vector $\boldsymbol{\mu}$ of means of p variables can be obtained by calculating the arithmetic mean of each column of the matrix and then assembling the results into a column or row vector, taking care to maintain the order of results.

Let's consider the data shown in the following table, where m_1, m_2 and m_3 are the observation taken on p_1, p_2 and p_3 dimensions of the variable x, and $\bar{x}$ is the average:

	p_1	p_2	p_3
m_1	2	4	8
m_2	3	5	3
m_3	4	6	1
sum	9	15	12
μ	3	5	4

The column-vector and the row-vector are thus the following:

$$\boldsymbol{\mu} = \begin{bmatrix} 3 \\ 5 \\ 4 \end{bmatrix} ; \ \boldsymbol{\mu}' = [3 \quad 5 \quad 4]$$

Alternatively, $\boldsymbol{\mu}$ can be computed directly using matrix algebra from the matrix of the m observations; thus, if **A** = m x p, then the formula is:

$$\boldsymbol{\mu}' = \mathbf{1}' \times \mathbf{A} \times (\mathbf{1}' \times \mathbf{1})^{-1} = (1 \times \mathrm{m}) \times (\mathrm{m} \times \mathrm{p}) \times ((1 \times \mathrm{m}) \times (\mathrm{m} \times 1))^{-1}$$

Here, **1** and **1'** are vectors consisting of a row and a column of ones, respectively, as many as m and p in **A**. Let's calculate the first step:

$$\mathbf{1}' \times \mathbf{A} = [1 \quad 1 \quad 1] * \begin{bmatrix} 2 & 4 & 8 \\ 3 & 5 & 3 \\ 4 & 6 & 1 \end{bmatrix}$$

$$= [(2*1+3*1+4*1) \quad (4*1+5*1+6*1) \quad (8*1+3*1+1*1)]$$

This step allows the calculation of the total for each column:

$$\mathbf{1}' \times \mathbf{A} = [1 \quad 1 \quad 1] \begin{bmatrix} 2 & 4 & 8 \\ 3 & 5 & 3 \\ 4 & 6 & 1 \end{bmatrix} = [9 \quad 15 \quad 12]$$

The inverse of the product $((1 \times \mathrm{m}) \times (\mathrm{m} \times 1))^{-1}$ results in the scalar 1/n used in the arithmetic mean calculation of n elements. This second step yields the following result:

$$(\mathbf{1}' \times \mathbf{1})^{-1} = \left([1 \quad 1 \quad 1] \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} \right)^{-1} = 0.3\overline{3}$$

Thus, putting together the two steps we get the following multiplication of a row-vector and a scalar:

$$\boldsymbol{\mu}' = \mathbf{1}' \times \mathbf{A} \times (\mathbf{1}' \times \mathbf{1})^{-1} = [9 \quad 15 \quad 12] * 0.3\overline{3} = [3 \quad 5 \quad 4]$$

Obviously, in order to obtain the row-vector as column-vector:

$$(\boldsymbol{\mu}')' = \boldsymbol{\mu}$$

The vector $\boldsymbol{\mu}$ has a single column because each mean of the p variables is treated as a single observation of the "average" variable, whereas $\boldsymbol{\mu}'$ has a single row as it represents the average for each of the p variable. Therefore, we shall call $\boldsymbol{\mu}$ the vector of averages, and $\boldsymbol{\mu}'$ the mean vector. Of course this only a mere distinction.

This note was adapted from: Berman HB. How to Compute Vector Means. Available from: https://stattrek.com/matrix-algebra/vector-mean. Accessed January 25th 2025.

Another illuminating contribution for the calculation of the variance-covariance matrix is the following: Berman HB. Variance-Covariance Matrix. Available from: https://stattrek.com/matrix-algebra/covariance-matrix. Accessed January 25th 2025.

NOTE 4 – Autocorrelation within the MLQC and correlation between QCs

Since an assumption of using the multivariate chart is the correlation between variables, in the case of the analytical process, it is necessary to demonstrate that there is a correlation between the levels of MLQC. However, one could argue that this correlation is actually produced by the decomposition of autocorrelated data, such as those in the time series in which the quality control results are generated by the instrument.

Autocorrelation is a phenomenon where an ordered series, such as a time sequence, has a structure in which each value is related to certain of its previous values. It is evident that if MLQC data are natively autocorrelated and not correlated in its level components, then using the multivariate control chart is no more appropriate than using a unique univariate chart for all the (properly standardized) data.

To demonstrate autocorrelation in the data series, it is important to remember that the MLQC is composed of three surrogate plasma samples (QC1, QC2, QC3), each corresponding to different levels of analyte concentration (low, medium, high). The instrument used to analyze the MLQC in this study uses a sample loader with fixed positions for the control levels. Therefore, even though the analyzer itself is a “random access” machine, the sampling of levels always occurs in the same order. In this study, the sample order is QC1-QC2-QC3, corresponding to the sequence of levels “low-medium-high.”

Considering that the analytical and non-analytical times of the instrument are roughly constant, this sampling scheme creates periodicity in the MLQC data. Indeed, if the data are observed as a whole in the order they are generated by the machine, without distinguishing the series by the level of MLQC they represent, a repetition of triplets of values can be observed, each corresponding to an analytical run. The

sinusoidal trend manifests itself due to the differences in concentration within the triplet, which is always repeated in the same order and with roughly the same magnitude.

It should be noted that the time series of all MLQC data has uneven spacing. In fact, the distance between data points within a triplet is significantly smaller (*i.e.*, minutes) than that between triplets (*i.e.*, days):

To capture this structure, two types of autocorrelation analysis are needed, which generate a characteristic function between the distance of autocorrelated elements and the size of the correlation:

1. Overall: simply called the autocorrelation function (ACF), where the correlation effect between each value in the series and k of its predecessors is collectively measured.
2. Partial: where the corresponding function (PACF) isolates the correlation effect of the k-th preceding element on the one of interest, excluding all intermediaries.

To analyze both types of autocorrelation, the original sequence needs to be correlated with itself but with a shift that aligns it with the k-th element. In other words, considering a sequence “a-b-c-d-e”, if k=1, the shifted sequence is “b-c-d-e,” and the correlation is done by pairing “a-b,” “b-c,” “c-d,” and so on, where the first element of the pair belongs to the original series and the second to the shifted series. Autocorrelation thus always concerns m-k elements, where m is the length of the

original sequence. In time series, the shift is called lag, and k is measured as the number of lags with respect to which ACF and PACF are calculated.

Since the series is composed of triplets of values, and our interest is to demonstrate that they are connected, the time series of data must predominantly be analyzed using the PACF. This is better understood by considering a prototype sequence of 3 triplets (corresponding to 3 analytical runs) QC1-QC2-QC3, which produces the following alignments up to k=6, *i.e.*, up to the autocorrelation between the first element of the leading triplet in the sequence ($QC1_1$) and the first element of the trailing triplet ($QC1_3$).

Table A. Lagged alignments of MLQC levels for autocorrelation analysis (subscript indicates the triplet)

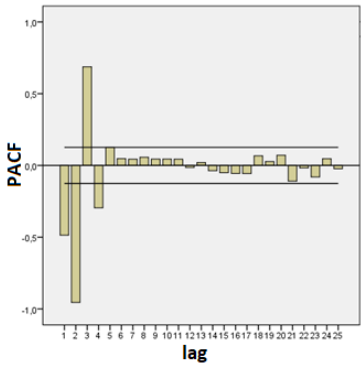
lag	*sequences*								
1	$QC1_1$	$QC2_1$	$QC3_1$	$QC1_2$	$QC2_2$	$QC3_2$	$QC1_3$	$QC2_3$	$QC3_3$
	$QC2_1$	$QC3_1$	$QC1_2$	$QC2_2$	$QC3_2$	$QC1_3$	$QC2_3$	$QC3_3$	-
2	$QC1_1$	$QC2_1$	$QC3_1$	$QC1_2$	$QC2_2$	$QC3_2$	$QC1_3$	$QC2_3$	$QC3_3$
	$QC3_1$	$QC1_2$	$QC2_2$	$QC3_2$	$QC1_3$	$QC2_3$	$QC3_3$	-	-
3	$QC1_1$	$QC2_1$	$QC3_1$	$QC1_2$	$QC2_2$	$QC3_2$	$QC1_3$	$QC2_3$	$QC3_3$
	$QC1_2$	$QC2_2$	$QC3_2$	$QC1_3$	$QC2_3$	$QC3_3$	-	-	-
4	$QC1_1$	$QC2_1$	$QC3_1$	$QC1_2$	$QC2_2$	$QC3_2$	$QC1_3$	$QC2_3$	$QC3_3$
	$QC2_2$	$QC3_2$	$QC1_3$	$QC2_3$	$QC3_3$	-	-	-	-
5	$QC1_1$	$QC2_1$	$QC3_1$	$QC1_2$	$QC2_2$	$QC3_2$	$QC1_3$	$QC2_3$	$QC3_3$
	$QC3_2$	$QC1_3$	$QC2_3$	$QC3_3$	-	-	-	-	-
6	$QC1_1$	$QC2_1$	$QC3_1$	$QC1_2$	$QC2_2$	$QC3_2$	$QC1_3$	$QC2_3$	$QC3_3$
	$QC1_3$	$QC2_3$	$QC3_3$	-	-	-	-	-	-

In contrast, breaking the series into 3 sub-series produces the alignments presented in Table B.

Table B. Correlation analysis of MLQC levels (subscript indicates the triplet)

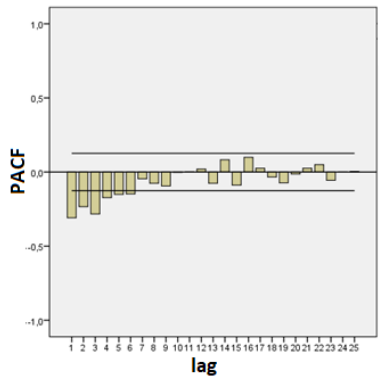
Bivariate correlation	*alignment*		
r(QC1, QC2)	$QC1_1$	$QC1_2$	$QC3_1$
	$QC2_1$	$QC2_2$	$QC3_2$
r(QC1, QC3)	$QC1_1$	$QC1_2$	$QC3_1$
	$QC3_1$	$QC3_2$	$QC3_2$
r(QC2, QC3)	$QC2_1$	$QC2_2$	$QC2_1$
	$QC3_1$	$QC3_2$	$QC3_2$

The PACF of the series (for m=84 triplets) is as follows (calculations were made with SPSS 20, IBM Company, Armonk, NY):

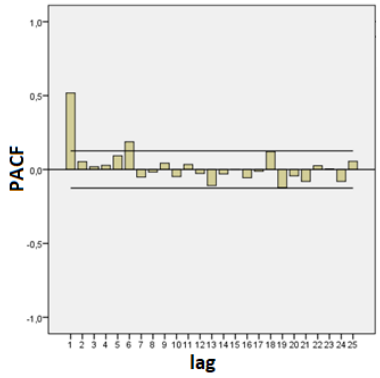


Partial autocorrelation is observed up to lag 4 (*i.e.*, k=4), and it peaks at lag 2. Looking at Table B, this is consistent with the previously mentioned “forced” periodicity of the data. To eliminate this masking effect, it is possible to rearrange the order of levels within each triplet, keeping the order of triplets unchanged, and standardize the data values using the mean and SD specific to each level. This smooths out the disparities due to concentration level and focuses solely on variation

as such. The PACF obtained by randomizing one time the data within triplets is as follows:



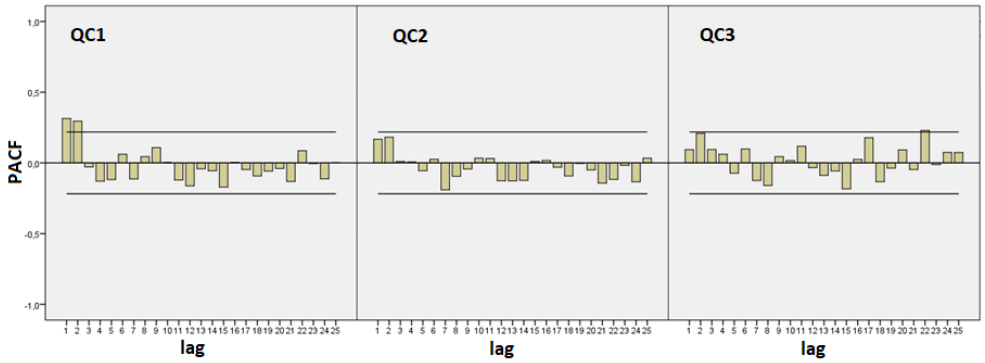
Thus, after standardizing, we find that:



Since one of the operations involved randomization, the single result could merely be due to chance and needs to be “stabilized” by repeating it a sufficient number of times, such as 100. If we consider autocorrelation at a certain lag to be unstable if it appears fewer than 5 times out of 100 (*i.e.*, $\alpha = 0.05$), then we find that the only stable autocorrelation is at lag 1.

Thus, in isolation, the data show chaining: QC1 correlates with QC2, QC2 with QC3, and QC3 with QC1 (of the next triplet), even when the order within the triplets is randomized and the data standardized. This suggests that the data within a triplet naturally move together, and this movement is only slightly affected by the partition into triplets (otherwise, we would have seen correlation at lag 3 or higher).

Regarding this last point, it is interesting to explain why, if the data are clustered into triplets, there is a correlation between QC3 and QC1 across triplets. Notably, the PACF of the series of the three levels in isolation shows:



QC1 exhibits autocorrelation at lag 1 and lag 2, unlike the other two series, which show none*. Since QC1 is autocorrelated up to lag 2, the QC1 of each triplet "resembles" that of the next two triplets.

Thus:

1. The lag 1 correlation of the MLQC is explained by the lag 1 and lag 2 autocorrelation of QC1, allowing QC3 to correlate with QC1 across consecutive triplets.
2. The absence of a lag 2 autocorrelation in the MLQC data sequence is explained by the absence of autocorrelation in the QC2 and QC3 sequences, with a mechanism analogous (but inverse) to point 1.

This suggests that the time series of MLQC with "peculiar" uneven spacing, results from the offset assembly of correlated series. Therefore, values move more coherently within triplets than between them. The coherence within the triplet shows they are the product form the same process, but on different levels. As such, the MLQC needs to be treated as a multivariate object and not as the assembly of univariate pieces.

[*]This different behavior is likely due to the fact that in a nonlinear calibration curve, as used by the analytical method that generated the data, the behavior in the low concentration range is much more “constrained” than at intermediate or high concentrations due to proximity to the quantification and detection limits, which are influenced by instrumental background noise. This limit is further raised by chemical noise from the progressive instability of reagents, further reducing the freedom of variation in results. This likely induces a memory effect at low concentrations, which is reflected in autocorrelation in the series.

NOTE 5 – A Microsoft® Excel spreadsheet for the analysis of multivariate data with Hotelling's T^2 control chart (with synthetic data generator)

Surely, most readers of this work, if not all, are familiar with constructing Shewhart univariate charts using electronic spreadsheets. This is possible because calculations of the standard deviation, mean, and percentiles of the normal distribution are implemented as basic spreadsheet functions, requiring knowledge of elementary.

Fortunately, the same spreadsheets now offer users the most common matrix algebra operations (addition, subtraction, multiplication, inversion, and transposition) and percentiles of many probability distributions that are used in the creation of control charts. In Microsoft® Excel these functions are the following ones:

- MMUMLT→ multiplication (see point "g" in Note 1)
- MINVERSE → inverse (see point "i" in Note 1)
- TRANSPOSE → transposition (see point "j" in Note 1)

The spreadsheet provided as supplementary material for this work, and described in this Note, allows for three fundamental activities regarding multivariate quality analysis:

- Establishing the T^2 chart with historical data in Phase I
- Analyzing multivariate Phase II data using the chart created in Phase I
- Generating correlated data for free manipulation and simulating Phases I and II

An additional activity includes the ability to calculate an extra control level based on APS (Acceptable Performance Standard), by entering the values of the maximum acceptable deviation for each level of MLQC (Multivariate Linear Quality Control).

The content, function, and use of the various worksheets in the spreadsheet are described in the INSTRUCTIONS sheet.

It should be noted that to comply with what has been stated so far, the spreadsheet does not use macros but only functions that can be called via cell syntax. Therefore, all calculations otherwise unavailable are explicitly executed by using different cells within the same sheet to develop the necessary components and steps to complete the task. This can be clearly seen in the case of the covariance matrix with the Holmes-Mergen estimator (see the COV MATRIX sheet), as well as in the generation of correlated data with the Cholesky decomposition (see the CORELATED DATA GENERATOR sheet).

Obviously, by not using macros, the spreadsheet cannot dynamically adapt to the size of the data sample entered. Therefore, only data triplets can be used - necessarily 30 for constructing the chart and up to 12 for analysis. This choice is intended to encourage readers to replicate the processes described in this work, inviting them to subsequently use this same spreadsheet with their own data.