AUTOMATIC PROCESSING OF LARGE BATCHES OF EXPERIMENTAL DATA FOR THE INITIAL CALIBRATION OF LHC CRYOGENIC THERMOMETERS[*]

F.Pavese[1], D.Ichim[1], P.Ciarlini[2], C.Balle[3], J.Casas-Cubillos[3]

[1]CNR, Istituto di Metrologia "G.Colonnetti", Torino, Italy
[2]CNR, Istituto per le Applicazioni del Calcolo "M.Picone", Roma, Italy
[3]CERN, Geneva, Switzerland

ABSTRACT

In modern times, measurements are performed using computers and when the number of measurements and/or of the resulting data is large, automatic methods must be used in all stages, from data acquisition to data reduction and analysis. Methods have been developed by IMGC under a CERN Contract for give a sound mathematical and statistical basis to the operations needed during all the above steps, in connection to the needs of the calibration of the cryogenic thermometers for the LHC. However, these methods have a wider generality and can be useful in many other experimental instances. The paper introduces methods for optimisation of experimental design, for raw-data reduction in the presence of a drift in time and for detection of data clustering. The main results will be illustrated from a large initial number of calibration data.

1. INTRODUCTION

At the European Center for Nuclear Research (CERN) a new particle accelerator will be constructed, called “Large Hadron Collider” (LHC). For the construction of the LHC, about 6000 cryogenic semiconducting-type resistance thermometers will be individually calibrated with automatic equipment and data reduction procedures in the temperature range (1,6–300) K. The essential feature of the data set is the limited dispersion of the thermometer electrical-resistance vs. temperature characteristics. The automatic data handling concerns every step of the process: a) data acquisition from all the thermometers in a run at each calibration temperature T_i ; b) acquired data reduction to obtain, for each j-th thermometer, $R_{ji}(T_i)$; c) fitting the calibration points to obtain the parameters of the model function $T = \mathrm{f}(R,\boldsymbol{a})$ and detection of anomalies in the calibrations. The studies performed at IMGC include all these steps and will be introduced in the paper.

Concerning data acquisition, an acquired sample is generally made up of several instrumental readings. These readings are reduced to a single value by simple methods, usually by averaging. To avoid ambiguity in the results or a computation time too long with respect to the experimental constraints, the automatic data acquisition has to be as much as possible robust against outlying values. The paper first introduces an algorithm, named “sequence-analysis outliers rejection” (SAOR) that manages *on-line* the most usual problems affecting the measurand during the acquisition (non-linear thermal drift, outliers due to noise peaks) without resorting to complex statistical computations. The algorithm uses the ordering of the reading sequence and of the “distances” between successive readings. Results on tests performed for the equispaced case are reported using simulated data [1]. Concerning the subsequent steps, a computational method was used, the Least Squares Method with Fixed Effect (LSMFE), that takes into account both the physical similarity of the thermometers and their individuality [2]. Using it, the problem of the compensation for the thermal drift occurring during acquisition was efficiently solved, robust against the occurrence of outliers. It was subsequently used for detection of clusters of thermometers with inherently different characteristics. Finally, a study to

[*] Work partially funded by CERN (# 90562GCH12Q121)

identify the optimal calibration-point distribution was also developed, used for the data acquisition step.

2. A PRE-PROCESSING ALGORITHM FOR ON-LINE OUTLIER REJECTION BY SEQUENCE-ANALYSIS IN DATA ACQUISITION (*SAOR*)

The characteristics of the instrumental readings shown in Fig.1 can be summarised as follows:

- data acquisition consists in an ordered sequence of readings from an instrument (e.g., a digital voltmeter, bridge, …);
- the signal (e.g., resistance or voltage values) is assumed to be quasi-stationary during a sequence of readings;
- changes in signal value are due to three reasons, of different and generally independent origin:
  A) **signal drift** (reason: thermal drift): it is assumed to have a frequency spectrum limited to sufficiently low frequencies to allow modelling with a low-order polynomial;
  B) **signal noise** (reason: electrical instrumental noise: random variations of the readings with a broad frequency band -typically white noise): assumed to be *stationary* within each sequence of readings
  C) **pulse noise** (reason: spot events in the environment, of magnetic, electric (switching), mechanical (shocks), etc. nature): it occurs as random spikes assumed not to affect more than a maximum number *K* of consecutive readings (generally a few), but with no limits within the whole reading sequence as to the number of occurrences, the position within the sequence and the magnitude and sign.

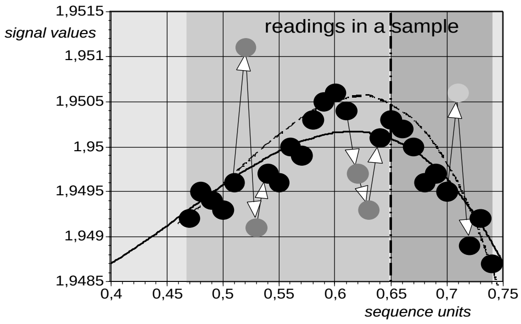

Fig.1 – A typical acquired sequence of instrumental readings (trend, due to signal drift, plus system noise). Noise peaks can also randomly occur (grey dots in the figure), affecting one or more consecutive readings. Signal drift is shown, fitted either with (solid curve) and without (broken curve) taking into account the noise peaks.

The SAOR algorithm does not compute a regression baseline (signal drift), but directly analyses the sequences of distances between consecutive instrumental readings. For readings non-uniformly spaced in time, the divided differences or the Euclidean distances should be adopted to take into account different scales in the two variables. Since equispaced data are the most common case in automatic acquisition, SAOR was presently developed for this case, where the vectorial distance between two consecutive readings simply is the projection on the *y*-axis. The algorithm inputs are: the number *N* of output data $\{(\tau_1 , \psi_1), \ldots, (\tau_N , \psi_N)\}$, the number $M > N$ of input initial instrumental readings (t_i , y_i), the assumed maximum number *K* of consecutive outlying *reading* values, and a threshold value d_o to discriminate possible outliers according to the knowledge of the signal characteristics. The algorithm steps are (the full treatment can be found in [1]):

1. acquire a number $M > N$ of instrumental readings in sequence: $\{(t_i , y_i)\}$, $i = 1,\ldots, M$;

2. compute the projection on the y axis of the (vectorial) distance between two consecutive readings. Associate to each $d_i = |y_{i+1} - y_i|$, $i = 1,\ldots, M$-1 and $s_i = \pm 1$ for $(y_{i+1} - y_i) \gtrless 0$. In the following d_i is called "distance" and the sign s_i is called "direction";
3. given a threshold d_0 for these *distances*, define as *candidate outliers* those distances for which $d_j > d_0$, and compute the number C of their occurrences;
4. if $M–C–1 < N$, acquire supplementary instrumental readings as necessary and return to step 2;
5. starting from *each candidate outlier*, say the j-th, analyse the sub-sequence of length $L = K$+1 for consecutive distances, d_j ,..., d_{j+K};
6. define as **not** an *outlier* a candidate distance, whenever the relative sub-sequence does **not** satisfy one of the following conditions:
   a) more than one candidate exists, except for the first and last distances;
   b) at least one change of the direction occurs.
7. for each one of the remaining sub-sequences, that do contain outlier readings, use a "truth table" of decisions to identify the outlying *readings* (from one to K). Clear off the outliers readings to output the cleaned sequence of readings (τ_k , ψ_k), $k = 1,\ldots,N$.

In step 3 the value of the threshold depends on the scale of the measures. To obtain the sample value and its uncertainty, simple average of the "cleaned" sequence or regression can then be applied using a suitable functional model of the baseline.

To evaluate the performance of the algorithm, simulation studies have been performed. A basic simulated sequence was built-up including a random noise component and a non-monotonic baseline drift affecting the reading values by a factor of about 2. Then, this basic sequence was automatically altered by including outlying elements, random in number, up to a maximum value $O_{\max}$, in position in the sequence, in relative size, $R = y_{\max}/y$, and in "sign". The values of the simulation parameters were: N = 18; $M = N$+2, $O_{\max}$ = 4 (maximum number allowed for the *candidate outliers* in a sub-sequence: $O_{\max} < (M/2 – 1)/2$), with $O_{\max} > K$).

The routine (presently implemented in FORTRAN 77) takes only a few microseconds to run on a modern PC. Tests of groups of 10 000 trials gave essentially the same results, therefore no extension of the tests above 60 000 random sequences was considered necessary. The number of algorithm failures, i.e., of mismatches between the pre-imposed outliers and the ones recognised by the algorithm, was tested as a function of the maximum outlier size relative to the signal value, R. When $O_{\max}$ matches the discrimination threshold K = 2, i.e., when $O_{\max}$ = 1, the efficiency of the algorithm is 100%. When $O_{\max}$ > 1, there is a non-zero statistical probability for the outliers within the sequence to form sub-sequences affecting more than two consecutive readings, violating the case-study assumptions of the algorithm. This is the main reason for a resulting ≈ 0.5 % inefficiency at $O_{\max}$ = 4 for outliers of size much larger than the signal (R = 15). Some inefficiency (< 0.1 %, for $O_{\max}$ = 2) also arises from sequences involving a few special cases not accounted in the simple truth table [1] adopted, such as outliers sequence starting from the first reading: they might be taken into account but the computational cost is too high in most experimental cases. When the R-value is lowered to approach the signal value, the inefficiency grows gradually for R = 3–2 to a value of about 3 % (for $O_{\max}$= 4). When $R \rightarrow 1$ the algorithm obviously breaks, since most of the readings are declared candidate outliers.

3. APPLICATION OF THE LSM WITH FIXED EFFECT

3.1 The mathematical model

The physical similarity of the thermometers allowed the choice of a common model to fit all the data series. However, a systematic bias between series must also be included to take into account the specificity of each thermometer. In the case of an additive bias characterising each data series, the general description is: $y_{k,}$ = *Common*($x_{k,}$) + *Specific*($x_{k,}$), for $k = 1,\ldots,S$ and $i = 1,\ldots,\ n_k$ where S is the total number of series and n_k is the number of measurements of the k-th series, resulting in $N = \sum_{k=1}^{S} n_k$ total number of measurements. Let us assume that a polynomial function is suitable for the *Common* part and that the *Specific* part is linear. Then the following model function holds:

$$g(x) = \sum_{h=0}^{m} a_h x^h + \sum_{k=1}^{S} \delta_{j,k}\left(t_k + u_k x\right) \quad \text{where}\ \delta_{j,k} = \begin{cases} 1, \text{if } x \text{ in to the } k-\text{th serie} \\ 0, \text{otherwise} \end{cases} \qquad j = 1,\ldots,S \tag{1}$$

The M = m+2S+1 parameters of the previous equation are not independent. Hence the parameters a_0 and t_k are not independent, and, similarly, the parameters a_1 and u_k. The least square estimates of the following model equation with m+2S-1 independent parameters can be obtained, requiring $\sum_{k=1}^{S} t_k = 0$:

$$g(x) = \left(a_0 + t_f\right) + \left(a_1 + u_f\right)x + \sum_{h=2}^{m} a_h x^h + \sum_{k=1(-f)}^{S} \delta_{j,k}\left[\left(t_k - t_f\right) + \left(u_k - u_f\right)x\right] \tag{2}$$

where the f-th series is taken as reference and is skipped in the second summation.

Now the intercept $\delta = a_0 + t_f$ of the reference series also represents the mean value of the f-th series, while each new bias parameter $\tau_k = t_k - t_f$ represents an effect relative to the reference. Similar considerations are applied to the first derivative of the above equation for x = 0, by defining $\gamma = a_1 + u_f$ and $\upsilon_k = u_k - u_f$. Equation 2, assuming f = 1, gives the M-dimensional vector of the unknowns $\alpha = (\delta, \gamma, a_2, \ldots, a_m, \tau_2, \ldots, \tau_S, \upsilon_2, \ldots, \upsilon_S)$, the regression matrix $\left(\begin{array}{c|c} \boldsymbol{X} & \boldsymbol{0} \\ & \boldsymbol{D} \end{array}\right)$, where $\boldsymbol{X}$ is the (N, $m+1$) polynomial regression matrix, and the (N-n_1, 2(S-1)) matrix $\boldsymbol{D}$ = (d_{ij}), given by:

$$d_{ij} = \begin{cases} 1, \text{ if } j \le S \text{ and } x_i \text{ belongs to the } j+1-\text{th series} \\ x_i, \text{if } S \le j \le 2S-2 \text{ and } x_i \text{ belongs to the } j-S+2-\text{th series} \\ 0, \text{ otherwise} \end{cases} \tag{3}$$

3.2 Compensation of thermal drift

In the thermometer calibration equipment thermal drift in time always occurs. Therefore, using a method that allows obtaining the same accuracy without needing very stable thermal conditions, provided that an internal check can be made for detection of possible thermal gradients, can reduce the calibration time. For thermal drift correction, typically many measurements are added, taken from a reference calibrated thermometer. These additional data are fitted to obtain the drift function, to be subtracted from the measurements of the uncalibrated thermometers. The use of a procedure based on the LSMFE avoids the use of the reference thermometer, since it allows fitting the whole set of data acquired from the uncalibrated thermometers. Hence, a few data points (typically 4-6) for each thermometer, on a large number of thermometers (15–50), provides an accurate evaluation of the drift function with a better statistics, reducing the estimated variance [2]. The temperature span of the drift being limited, the influence of the effect of the thermometer individuality is limited too, and consequently the function *Specific* is a low-degree polynomial.

In any case, the presence of outlying data requires a preliminary screening procedure. These outliers can be due either to pulse-noise (section 2 [1]) or to outlying characteristics of a thermometer (either intrinsic or due to its mounting on the calibration apparatus). The outlier rejection for the second case has been simply obtained by adopting a threshold criterion.

The quality of the processed data series resulted to be quite variable, as shown in Fig.2 for low-drift cases, reflecting the run-to-run variability of the thermal conditions in the calibration apparatus, though generally matching the tolerated uncertainty.

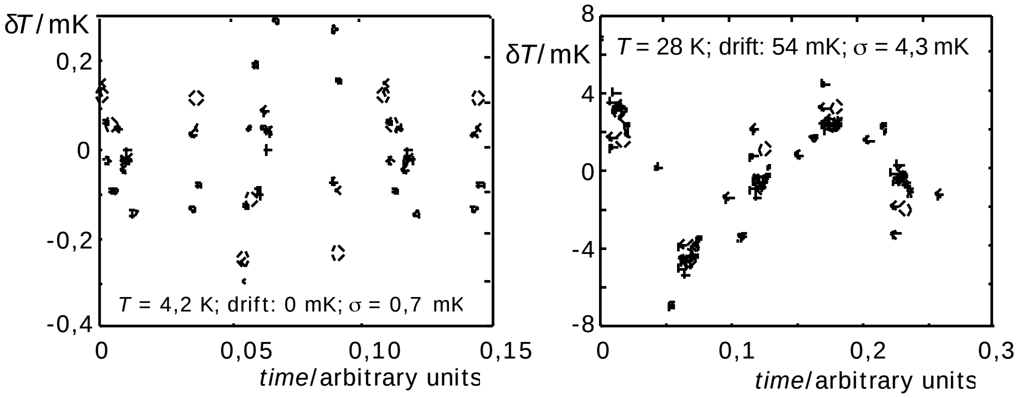

Fig.2 – Residuals of the overall fitting of 20 uncalibrated thermometers for low thermal drift at two temperatures. Different symbols indicate different thermometers. Above ≈ 30 K the total drift increases up to tenths of a kelvin. The σ values represent the actual uncertainty of the *whole* set of data.

When applying the procedure using the LSMFE to larger and larger acquisition time intervals by fitting more and more data acquired with higher drift rates, the dispersion of the residuals starts increasing in the high-drift region, indicating the start of a detectable thermal gradient between thermometers. This internal check can be used as a criterion for data selection.

3.3 Clustering analysis by fitting the *R*-*T* characteristics

In semiconductor thermometers, as the CERN LHC thermometers are, one possible problem to check is an excessive spread of the thermometer individuality (outlying characteristics) or the presence of clusters of thermometers with characteristics that require a different modeling. The procedure based on the LSMFE allows performing this check.

First, the *R*-*T* characteristics was obtained of the whole batch of thermometers using a logarithmic polynomial of degree *m*, with *m* = 4 for the *Common* part of the model and *m* ≤ 2 for the *Specific* part. When a plot of the residuals of the fit is done, as shown in Fig.3, it is possible to detect groups of thermometers with characteristics intrinsically different. This can be correlated with fabrication parameters (position of the sensing elements on the chip for with semiconducting thermometers) or with calibration parameters (mounting on the comparison block, overheating of the sensing element due to the measuring current).

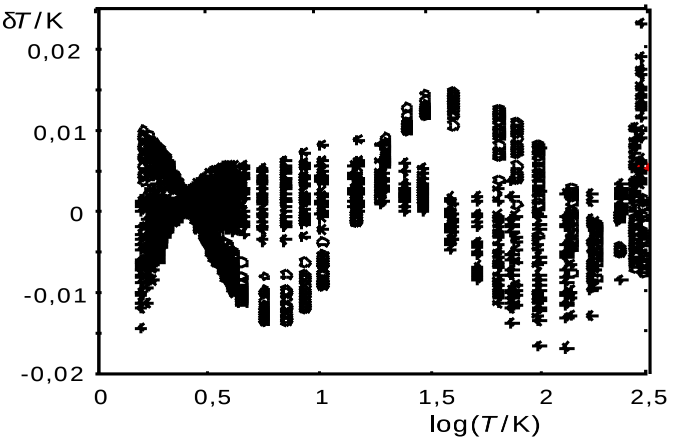

Fig.3 – Clustering of 80 thermometers. With the model used, two obvious groups of thermometers with different characteristics shows up in the fitting residuals. If the model is improved by increasing the order of the polynomial, two sets of values for the obtained parameters will show up, for the same two groups of thermometers and the residuals become random.

4. OPTIMISATION OF THE EXPERIMENTAL DESIGN

In the case of the CERN thermometers, the optimisation studies were performed on an initial sample of some hundred thermometers, each having been measured at 50 calibration points. From this data the study has determined the optimal temperature values at which the electrical resistance of the bulk of the thermometers will be measured to approximate the *R*(*T*) characteristics within given specifications. First, the choice of a suitable model sufficient to describe the *R*(*T*) characteristics was studied. Two models were considered (polynomials and splines) and a suitable transformation of the variables in order to obtain a better conditioning of the regression matrix [3]. The polynomial model was considered since it is widely used in metrology. It allows for an overall smoothing effect at the experimental points. In addition, the well-known numerical instabilities of the parameters were observed when increasing the order (up to the 9th degree), as well as oscillations in the derivatives. However, the local features of this type of thermometer also give rise to non-random residuals in the range where these features occur.

Using the spline model, on the contrary, which is known to be the most adequate model for local fittings, a sufficient model was found that captures all the important features of the *R*(*T*) characteristics, including their monotonicity. The quadratic splines proved not to be sufficient for the thermometer characteristics; considering that greater the spline order, greater their support is, the cubic splines were found sufficient. In the choice of the number of breakpoints, both the simplicity of the model and the quality of the fitting were taken into account. More breakpoints, higher the fitting quality is, but more complicated is the model. In order to match the local requirements of the thermometer characteristics, the knots were distributed according to the fourth order divided differences. Using these strategies, the set of knots was limited to 13 (one of which being a double one) that sufficiently describes about 94% of the initial sample of thermometers. Then, two measurement points (temperature values) were uniformly distributed within each interval, resulting in a total of 26 experimental points. As a final check, it was verified backward that this set is sufficient for approximating the curve obtained by fitting the initial 50 observations.

5. CONCLUSIONS

A group of procedures have been introduced, implemented in software routines (FORTRAN 77 for SAOR and MATLAB® for the LSMFE and splines, available on request) that can efficiently be used at different stages of the data acquisition and processing of calibrations of classes of thermometers with limited in-homogeneity [4]. They proved to be robust against outlying data or to be able to detect different classes of thermometers or to study the effect of experimental design on the whole batch of calibrations. Studies are planned to continue on the statistical evaluation of the results by using non-parametric methods (e.g., bootstrap).

REFERENCES

1. Pavese F., Ichim D., Ciarlini P., in *AMCTM V*, World Scientific, Singapore, 2001, 282-290.
2. Ciarlini P. and Pavese F., *Numerical Algorithms* 1993, **5**, 479-489
3. Ichim D., Ciarlini P., Pavese F., *Optimisation of Data Point Distribution Fitting for the LHC CERN Thermometers by Spline*, IMGC Technical Report N°45 (in english), Torino, March 2001.
4. CERN Report, in preparation.

Addresses of the Authors: Dr. F.Pavese and Dr. D.Ichim, CNR, Istituto di Metrologia “G.Colonnetti” (IMGC), strada delle Cacce 73, 10135 Torino, Italia. E-mail: f.pavese@imgc.to.cnr.it, d.ichim@imgc.to.cnr.it; Dr. P.Ciarlini, CNR, Istituto per le Applicazioni del Calcolo “M.Picone) (IAC), viale del Policlinico 173, 00161 Roma, Italia. E-mail: ciarlini@iac.rm.cnr.it; Dr. C.Balle and Dr. J.Casas-Cubillos, CERN, Geneva, Switzerland. E-mail: c.balle@cern.ch, j.casas-cubillos@cern.ch.