XVIII IMEKO WORLD CONGRESS
Metrology for a Sustainable Development
September, 17 – 22, 2006, Rio de Janeiro, Brazil

REFERENCE VALUE IN HIGH FREQUENCY POWER COMPARISONS

Luciano Brunetti [2], Luca Oberto [1,2], Marco Sellone [1,2]

[1] Politecnico di Torino, corso Duca degli Abruzzi 24 – 10129 Torino, Italy, oberto@inrim.it, sellone@inrim.it
[2] Istituto Nazionale di Ricerca Metrologica, Strada delle Cacce 91 – 10135 Torino, Italy, brunetti@inrim.it

Abstract: According to the modern trend, the reference value of a measurement comparison among laboratories is established considering the contribution of all the participants appropriately. The main problem is deciding whether the data are consistent or they have to be discarded because of the evidence that the measured value is too different from the expected one. In this paper, the problem of the data rejection is analyzed for a specific comparison concerning microwave power measurements and a decision algorithm is presented.

Keywords: reference value, comparison, microwave power.

1. INTRODUCTION

In interlaboratory comparisons it is necessary to have a *Reference Value* with low uncertainty but the most important thing is the reliability of the value itself. Generally, the reference value is provided by a laboratory that can claim measurement uncertainties significantly lower than that of the participants. This is, usually, the pilot laboratory and very often is the National Primary Laboratory. In some cases this is not possible, so a *Consensus Value*, i.e. the best estimate of the measurand, has to be derived from the values produced by the same participants to the comparison.

The Consensus Value has a low uncertainty but it can be strongly biased by the inclusion, in the calculations, of anomalous data provided by some laboratory which made involuntary mistakes.

Different techniques can be used for the determination of the consensus value. In this paper we analyze the application of an algorithm that turned out to be less critical and more efficient then others normally used in the specific field of the microwave power measurements.

2. DATA REJECTION

The problem of the data refusal is quite spiny [2]. Sometimes it happens that one or more measurements seem to be in grinding discord with all the others. In this condition, the operator has to decide whether data are anomalous or can be used *bona fide*.

Unfortunately, it is not always possible to find the external cause of an anomalous result, so one must decide if has to discard the value or not. The decision is, in last analysis, subjective and one must use the maximum intellectual honesty in order to avoid prefixing its results.

In presence of data suspected of being anomalous in a relatively small set, the only honest behavior is to repeat the measures in order to identify the problem. When it is not possible to follow this way, we need some criterion to reject unreliable contribution.

One of the most used decision rule is the *Chauvenet criterion*, which applies to Gaussian distribution [2]. Having a suspected measurement in a particular data set, the Chauvenet criterion gives the probability to find other measures that differs, from the reliable ones, by a quantity comparable to the deviation associated to the suspected measure itself. Having large data sets, another simple and fast algorithm could be more useful for an efficient numerical computation. In this paper different solutions are examined for a specific data set obtained in a real exercise.

3. THE POWER MEASUREMENT COMPARISON SIT.AF-01 AT MICROWAVES

Before the theory, we introduce the experimental exercise to which they have been applied.

The INRiM (Istituto Nazionale di Ricerca Metrologica, formerly IEN "Galileo Ferraris", Torino, Italy) promoted a national power comparison in the microwave region (50 MHz–26.5 GHz) aimed to determine whether the laboratories accredited by SIT (Servizio Italiano di Taratura) for the *microwave power* quantity operate within their claimed uncertainty. In this particular exercise, a power meter with two coaxial sensors has been sent, as a traveling standard, to all the participants. Their task was to measure both the power sensor calibration factors K at fixed frequencies and the 1 mW - 50 MHz reference source included in the traveling standard. In other words, the comparison was an exercise of absolute power and power ratio measurements.

The pilot laboratory (INRiM) had to circulate two coaxial power sensors one of which fitted with 7 mm type N connector and the other with 3.5 mm connector to cover the mentioned frequency band.

Even though these sensors are traceable to the primary power standard, i.e. the microcalorimeter [3], they have to be calibrated with a routine method that cannot provide the best accuracy available. The reason of the choice to send this kind of sensors was in the technical impossibility to circulate bolometric detectors at that time, which conversely could be calibrated directly with the microcalorimetric technique, obtaining the best uncertainty allowed by the actual state of the art.

Because the pilot laboratory was not able to provide a Reference Value with a significantly small uncertainty, this was typical case in which a Consensus Value had to be derived from the data provided by all participants.

The first problem in this exercise was to find the unreliable measurements, if any. These are due to two different causes: a huge error that makes the measured quantity strongly different from the expected one or an underestimation of the measurement uncertainty that can cause an incorrect attribution of high reliability of the data itself.

4. THE SELECTION ALGORITHMS

The quality of a inter-laboratory comparison depends on the ability of distinguish between reliable measurements and unreliable ones. The better is the reference value, the lower is the uncertainty of the test.

In this section suitable algorithms are presented for the mentioned purpose, following a scheme also given in a specific references [4].

4.1. The median algorithm

This algorithm defines the median of all the measured results. The median is the middle element of a distribution: half of the results is above the median and half is below. All the results that contains the median in their uncertainty range are considered reliable and used in the average process for the determination of the consensus value, the other are considered unreliable and discarded.

The algorithm works well in eliminating results with unrealistically low uncertainty but fails if the data set includes only few results with low but realistic uncertainty among a majority of results with significantly higher uncertainty.

4.2. The cumulative probability algorithm

The median algorithm assigns, in the determination of the exclusion value (the first estimate of the reference value, i.e. the median) the same weight to all the results.

It is quite obvious that, to obtain a more robust algorithm, is necessary to assign different weights to the results: values with lower but realistic uncertainty must have a higher weight while lower weights have to be assigned to higher uncertainty values. This can be done considering each measured value belonging to a normal distribution with a standard deviation equal to one half of the declared uncertainty. This is consistent with the assumptions of the ISO Guide *GUM* [5]. The cumulative distribution of all the measurements is then calculated from an average of the single gaussian distributions [4].

This algorithm resolves the problems associated to the median algorithm but still it relies very heavily on the assumption that all the laboratories return correct values and associated uncertainties. This is the hypothesis the exercise has to validate in our case, anyway. The drawback of this algorithm is that an outlier result with very low uncertainty can be "overpowered" so much to polarize the exclusion value.

4.3. The "Value Voted Most Likely to be Correct" algotithm

The examination of the previous algorithms leads to the conclusion that it is important both to assign a weight to the uncertainty of the data and also do not overestimate such weights.

The *Value Voted Most Likely to be Correct* (VVMLC) algorithm interprets the uncertainty range of each participant as a rectangular distribution instead of a Gaussian one. The distributions are modified in such a way that the heights are one regardless of their widths. The name of the algorithm comes from the observation that this way of considering the distribution is like saying that each participant gives basically one vote to each value within its uncertainty range and no votes for values outside this range.

The cumulative distribution is determined by tallying the votes: one can determine, as the value (or range of values) considered likely to be correct, the value that receives the highest number of votes from the participants. This value becomes the first estimate of the Consensus Value of the exercise. If a range of values with equal (maximal) probability is found, the Consensus Value is chosen as the central value of the range.

Results cited in [4] show that the VVMLC algorithm is not only more robust than the median algorithm but also, in most cases, identify more participant values as reliable.

The way of tallying the votes prevents the overestimation of the weight for the low uncertainty values also. So it should give a more realistic Consensus Value.

5. DETERMINATION OF THE CONSENSUS VALUE

Once determined the reliable data set, the Consensus Value (*CV*) and its uncertainty (σ_{CV}) are calculated through a weighted mean of the values coming from the participant considered reliable:

$$CV = \frac{\sum_{i=1}^{n} \frac{m_i}{\sigma_{m_i}^2}}{\sum_{i=1}^{n} \frac{1}{\sigma_{m_i}^2}} \qquad \sigma_{CV} = \sqrt{\frac{1}{\sum_{i=1}^{n} \frac{1}{\sigma_{m_i}^2}}} \tag{1}$$

where n is the number of reliable participants, m_i are the measured values and σ_{mi} are the corresponding uncertainties.

6. APPLICATION TO THE MICROWAVE POWER COMPARISON SIT.AF-01

The VVMLC algorithm has been applied to the microwave power measurement exercise SIT.AF-01 in order to find the SIT laboratories needing a revision of their measurement techniques and methods, if any.

A Mathematica [6] code has been written that receives as inputs the values measured by the participants along with their claimed uncertainties. The computational code outputs the value of the first estimate of the Consensus Value, a plot of the cumulative distribution, the selection between reliable and unreliable data and the final Consensus Value and uncertainty as defined in (1). This result was additionally compared, by means of a compatibility test, with the data obtained at the INRiM High Frequency Laboratory before and at the end of the circulation, in order to be sure that the process is in agreement with the primary power standard.

Comparing data of participants and Consensus Values, it results that one laboratory gave completely unreliable data at all the tested frequencies. A detailed analysis of this case evidenced a systematic procedure mistake. Some other laboratories provided results a little bit lower or higher than the expected, at least at some frequencies. This evidence has to be carefully evaluated in order to find which problem affected these results.

It is important to note that some laboratories are traceable to foreign primary laboratory. This exercise proved that, as expected, different laboratories traceable to different primary standards are able to provide consistent data, except for little differences at some frequencies that has to be further investigated.

Fig. 1 shows the results collected for power sensor HP8481A (7 mm coaxial line transfer standard with type N connector) at the measurement frequency of 1 GHz. The straight line represents the final Consensus Value K_{ref} and the dashed lines represents its extended uncertainty $U(K_{ref})$ ($k=2$).

It can be seen that laboratories provided good measurements that are, along with their declared extended uncertainties, clearly compatible with the calculated Consensus Value.

Only laboratory L11, which did procedure mistake, is not compatible.

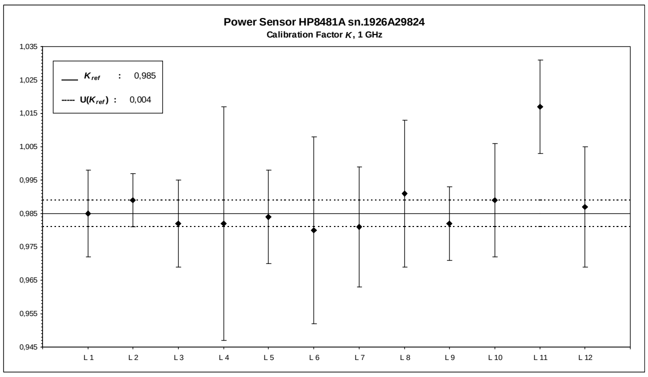

Fig. 1. Results of the participant laboratories, Consensus Value and related uncertainties for the calibration coefficient *K* of the 7 mm traveling standard at 1 GHz.

Fig. 2 shows the corresponding cumulative distribution obtained according to the VVMLC algorithm. The first estimate of the Consensus Value is the central value of the computed cumulative distribution highest peak. This value is used to verify if a measurement has to (or must not) be used for the determination of the final Consensus Value. This selection is done observing whether the first estimate of the Consensus Value lies in the one sigma declared uncertainty of a measurement. If so, the measurement is considered reliable, otherwise it is discarded.

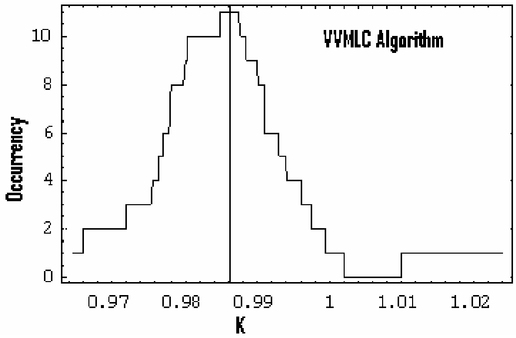

Fig. 2. Corresponding cumulative distribution obtained from the VVMLC algorithm.

Another meaningful example is presented in fig. 3. These are the measurements performed on the 7 mm transfer standard at the frequency of 10 GHz.

It can be seen that the general trend is very good except again for laboratory L11 as expected. Nevertheless laboratories L9 and L12 are not compatible at this frequency. In particular L9 underestimates the measurand and L12 is a little bit higher. L12 is not compatible even if has confidence band overlapping the confidence band of the Consensus Value. It is an example of little differences that need data rejection and further investigations.

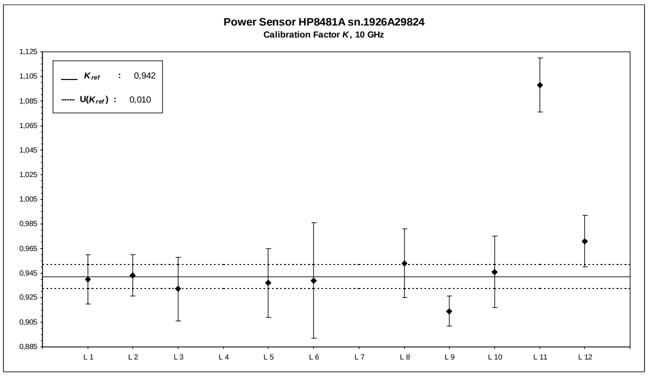

Fig. 3. Results of the participant laboratories, Consensus Value and related uncertainties for the calibration coefficient *K* measured on the 7 mm traveling standard at 10 GHz.

The compatibility of a measurement with the Consensus Value is determined by using the standard compatibility index test [7]:

$$-1 < \frac{K_{lab} - K_{ref}}{\sqrt{U(K_{lab})^2 + U(K_{ref})^2}} < 1 \qquad (2)$$

If relation (2) is satisfied, the measured value K_{lab} is compatible with the Consensus Value, otherwise it needs investigations. In this case the Consensus Value is denoted as K_{ref} because it is the value of the calibration factor K of the sensor that considered as the reference (consensus) value for the comparison (evaluated as in section 5).

The reason for which the L12 measurement cannot be accepted even if there is an overlapping between the two sigma confidence band of the Consensus Value resides in the choice of using the compatibility index test as the criterion for the measurements compatibility with the Consensus Value. Indeed this method is more precise than the simple overlapping – which corresponds to the request that the difference between two measurement is smaller than the sum of the two measurements uncertainties – because requires that the two uncertainty bars of the values under confrontation are not only overlapped but overlapped of a certain amount.

The choice of this method allows to be more confident about the goodness of the data provided by the laboratories participating in the circulation.

Fig. 4 shows the corresponding cumulative distribution obtained with the VVMLC algorithm. The first estimate of the Consensus Value is the vertical bar and the L11 outlier is clearly visible on the right (values about 1.1).

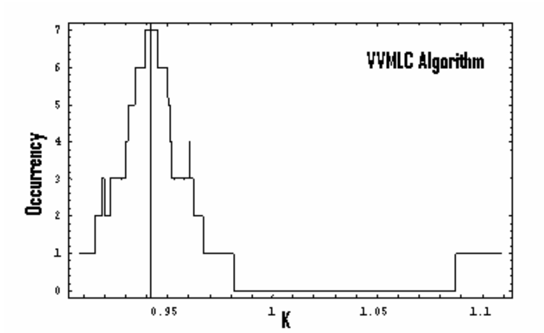

Fig. 4. Corresponding cumulative distribution (VVMLC algorithm).

In fig. 5 is reported another significant test, that is the measurement of the absolute power supplied by the reference source (P = 1 mW at 50 MHz) of the traveling power meter (a HP438A).

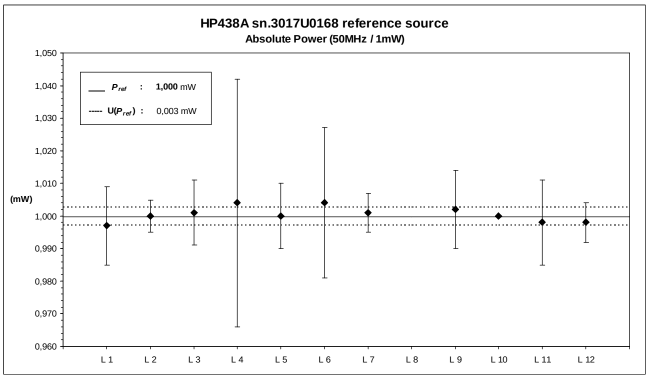

Fig. 5. Results for the absolute power measurement of the traveling standard reference source.

It can be seen that, in this particular measurement, all the laboratories have good performances including L11, which is unreliable in all the other cases. Laboratory L8 has not supplied its measurement while, in the L10 result, the uncertainty is missing.

Fig. 6 shows the corresponding cumulative distribution.

From these example we can say that the results obtained from the circulation are positive, the reason for the main discrepancies with laboratory L11 was found while some minor problems are currently under investigation.

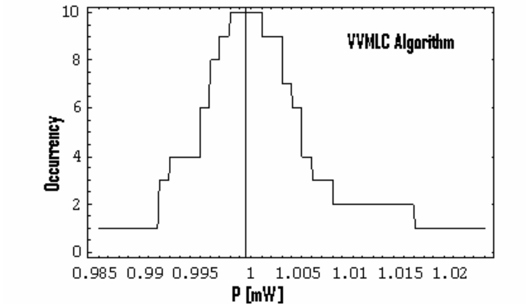

Fig. 4. Corresponding cumulative distribution (VVMLC algorithm).

7. CONCLUSION

The problem of the data rejection in interlaboratory comparisons is here analyzed and some examples of possible numerical algorithm for the determination of the Consensus Value are presented. The *Value Voted Most Likely to be Correct* (VVMLC) algorithm can be assumed as a good choice for the evaluation of the reliability of data and, therefore, it has been applied to the Italian Microwave Power Measurement Comparison exercise SIT.AF-01 with good success. Some results of this application are also presented and discussed. The VVMLC algorithm helped the pilot laboratory in the identification of the laboratories that needs a revision of their measurement technique. In particular one laboratory committed a mistake in the application of the measurement procedure that leads to completely unreliable data, some two other laboratories expressed little differences that needs further investigations.

ACKNOWLEDGMENTS

The authors wish to thank the SIT (Servizio Italiano di Taratura) for the support in the organization of the measurement exercise and all the participant laboratories.

REFERENCES

[1] L. Brunetti, L. Oberto and P. Terzi, "Confronto nazionale interlaboratorio SIT.AF-01. Potenza AF in linea di trasmissione coassiale da 50 MHz a 26,5 GHz", SIT Tachnical Report SIT.AF-01/06, March 2006.

[2] J. R. Taylor, "An introduction to error analysis. The study of uncertainties in physical measurements", University Science Books, 1982.

[3] L. Brunetti and E. Vremera, "A new microcalorimeter for measurements in 3.5mm coaxial line", IEEE Trans. Instr. Meas., Vol. 52, No. 2, pp. 320-323, April 2003.

[4] H. S. Nielsen, "Determining Consensus Value in interlaboratory comparison and proficiency testing", NCSLI Newsletter, Vol. 44, n. 2., April 2004.

[5] "Guide to the expression of the uncertainty in measurement", ISO, Geneva, 1993.

[6] Mathematica, http://www.wolfram.com.

[7] European co-operation for Accreditation, "EAL Interlaboratory Comparison (previously EAL-P7) Withdrawn", EA-2/03 rev. 1