Measurement 42 (2009) 1318–1323

Contents lists available at ScienceDirect

Measurement

journal homepage: www.elsevier.com/locate/measurement

Establishing reference value in high frequency power comparisons

Luciano Brunetti [a,*], Luca Oberto [a,b], Marco Sellone [a,b], Paolo Terzi [a]

[a] Istituto Nazionale di Ricerca Metrologica, Strada delle Cacce 91, 10135 Torino, Italy
[b] Politecnico di Torino, corso Duca degli Abruzzi 24, 10129 Torino, Italy

ARTICLE INFO

Article history:
Received 15 January 2008
Accepted 21 August 2008
Available online 30 August 2008

Keywords:
Reference value
Comparison
Microwave power

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 specific decision algorithm is presented.

1. Introduction

In interlaboratory comparisons it is necessary to have a *reference value* with low uncertainty but the most important aspect is that this value must be reliable, that is it must be as much as possible consistent with the expected one. The reference value is provided by the laboratory that can claim measurement uncertainties significantly lower than that of the other participants. Usually this is the pilot laboratory and very often this task is covered by the primary laboratory which can provide the best uncertainty level. However, in some cases this is not possible because, for example, the value provided by the pilot laboratory is obtained with technique and instrumentation that give an uncertainty comparable with that used by the other participants. So a *Consensus Value* (CV), i.e. the best estimate of the measurand, has to be derived from the values supplied by the same participants to the comparison.

The Consensus Value must have a low uncertainty, of course, but most important it has not to be strongly biased by anomalous data involuntarily provided by some laboratory. In order to minimize the biasing problem, it is necessary to find a way to identify and consequently reject the anomalous data from the calculation of the Consensus Value itself.

Different techniques can be used for the determination of the Consensus Value: in this paper we analyze the application to a national comparison exercise, concerning microwave power measurements [1], of an algorithm used for the data rejection, that turned out to be less critical and more efficient than others normally used in the specific field.

2. Data rejection

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

It is not always possible to find the external cause that explains an anomalous result and one must only decide if has to discard the value or not. In last analysis, the decision is subjective and one must use the maximum intellectual honesty in order to avoid prefixing his results.

In presence of data suspected of being anomalous in a relatively small set, the only honest solution is to repeat measurement in order to identify the problem. When it is not possible to follow this way because, for example, the

* Corresponding author. Tel.: +39 011 3919 323; fax: +39 011 3919 259.
E-mail address: l.brunetti@inrim.it (L. Brunetti).

0263-2241/$ - see front matter © 2008 Elsevier Ltd. All rights reserved.
doi:10.1016/j.measurement.2008.08.008

amount of data is huge or, as in the case of a measurement comparison, it is not possible to send again the standard to the laboratory that has made the mistake in order to investigate the problem, we need some criterion to reject unreliable contributions.

One of the most common decision rules 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 differ, from the reliable ones, by a quantity comparable to the deviation associated to the suspected measure itself. Another simple and fast algorithm could be more useful for an efficient numerical computation of the reference value both for the case of large data sets and for reduced number of data. In this paper different solutions are examined for a specific data set obtained in a real exercise able to return only a small number of data.

3. The power measurement comparison SIT.AF-01 at microwaves

Before explaining the used data rejection method, we introduce the experimental exercise to which it has been applied in order to focus the attention on a real problem. We chose an exercise with a reduced data set for seek of simplicity.

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 were operating within their claimed uncertainty or needed to modify their procedures. In this particular exercise [1], a power meter with two coaxial sensors has been sent, as a traveling standard, to all the participants. The traveling standard was identified in a Hewlett Packard power meter model 438A, chosen because it is a widely diffused and commonly used instrument in the laboratories involved. 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) circulated two coaxial power sensors one fitted with 7 mm type N connector and the other with 3.5 mm connector in order to cover the mentioned frequency band.

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

Since the pilot laboratory was not able to provide a reference value with an uncertainty significantly smaller than the other laboratories, this was a 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 if there were unreliable measurements given by the participants that had to be excluded from the computation of the Consensus Value. These unreliable measurements could be due to different causes: a huge error that makes the measured quantity strongly different from the expected one and that could be linked to mistakes made by the laboratory during the measurement process, or an underestimation of the measurement uncertainty that can cause an incorrect attribution of high reliability of the data itself.

4. The selection algorithm

The quality of an interlaboratory comparison depends on the ability of distinguishing between good measurements and unreliable ones. This ability allows obtaining a more reliable Consensus Value; the better is the reference value, the lower is the uncertainty of the test.

In this section suitable algorithms for the mentioned purpose are presented, following a scheme also given in a specific references [4]. The differences between them are highlighted in order to better explain our choice.

4.1. The median algorithm

The median is the middle element of a distribution in the sense that half of the results is above the median and half is below. To find the results to be discarded, the method considers the relation between the median and the measurement uncertainties. In particular, 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.

Since the algorithm evaluates the first estimate of the Consensus Value regardless to the uncertainties of the measured values, it 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 Consensus 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, it 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 (assuming that a coverage factor $k = 2$ has been adopted). 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, of giving a reliable weight to the data but it still 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. Another 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" algorithm

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

The *Value Voted Most Likely to be Correct* (VVMLC) [4] 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 so that any value contained in a distribution is taken into account one time. If the distribution associated to another result overlaps a region of values covered by the previous considered distribution, than the values contained in the overlapping region are taken into account two times and so on. The name of the algorithm comes from the observation that this way of considering the distribution is basically equivalent to saying that each participant gives 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 and 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.

Once found the first estimate of the Consensus Value, reliable results are selected in the same way as in the median algorithm.

Results cited in [4] show that the VVMLC algorithm is not only more robust than the median algorithm but also, in most cases, identifies 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 (u_{CV}) are calculated through a weighted mean of the values coming from the participants considered reliable:

$$CV = \frac{\sum_{i=1}^{n} \frac{m_i}{u_{m_i}^2}}{\sum_{i=1}^{n} \frac{1}{u_{m_i}^2}} \quad (1)$$

$$u_{CV} = \sqrt{\frac{1}{\sum_{i=1}^{n} \frac{1}{u_{m_i}^2}}} \quad (2)$$

where n is the number of reliable participants, m_i are the measured values and u_{m_i} are the corresponding uncertainties. The appropriate coverage factor has, then, to be applied to u_{CV}.

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 if there were SIT laboratories needing a revision of their measurement techniques and methods.

To analyze the data, 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), (2). 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, as expected, that 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 of the calibration factor K measurements collected for the power sensor HP8481A (7 mm coaxial line transfer standard with type N connector) at the 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, as already said, made procedure mistake, is not compatible. 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 can (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.

Another meaningful example is presented in Fig. 3. These are the measurements performed on the 7 mm

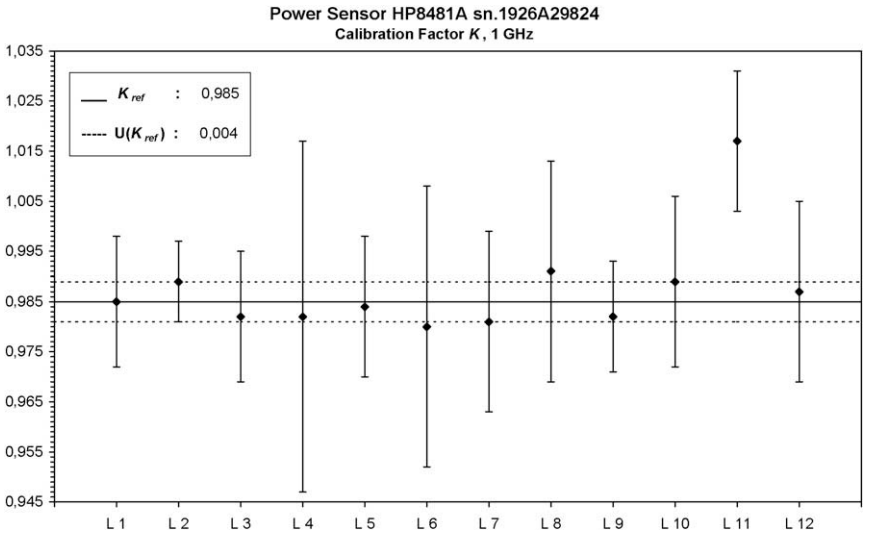

Fig. 1. Results of the participant laboratories, Consensus Value and related uncertainties for the calibration factor *K* of the 7 mm traveling standard at 1 GHz.

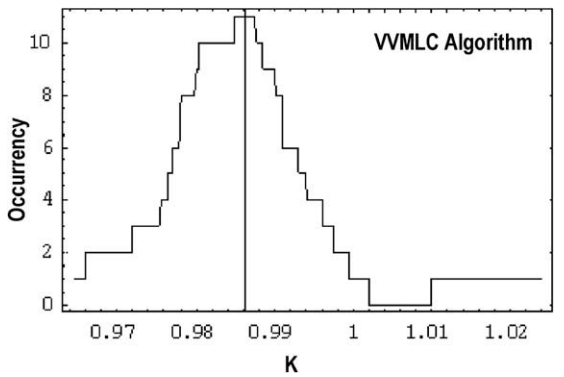

Fig. 2. Cumulative distribution obtained from the VVMLC algorithm for the calibration factor at 1 GHz.

transfer standard at the frequency of 10 GHz. It can be seen that the general trend is very good except again for laboratory L11. 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. This is an example of little differences that need data rejection and further investigations.

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

$$-1 < \frac{K_{\text{lab}} - K_{\text{ref}}}{\sqrt{U(K_{\text{lab}})^2 + U(K_{\text{ref}})^2}} < 1 \tag{3}$$

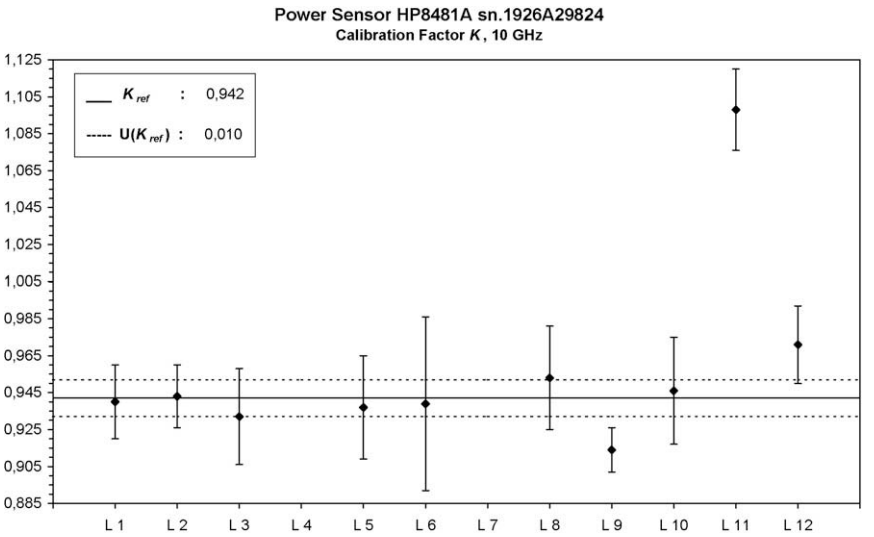

Fig. 3. Results of the participant laboratories, Consensus Value and related uncertainties for the calibration coefficient *K* measured on the 7 mm traveling standard at 10 GHz.

If relation (3) is satisfied, the measured value K_{lab} is compatible with the Consensus Value, otherwise it needs correction actions. 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 demanding than the simple overlapping – which corresponds to the request that the difference between two measurements 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, at least, 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).

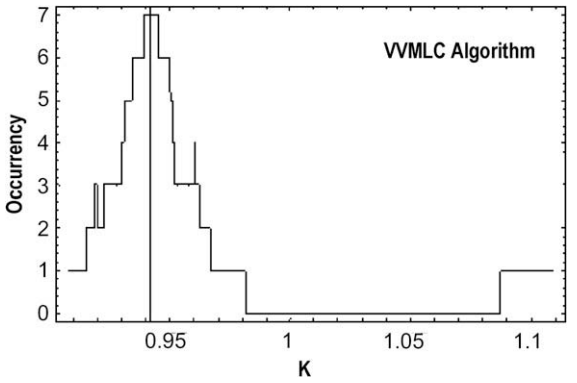

Fig. 4. Cumulative distribution (VVMLC algorithm) for the calibration factor at 10 GHz.

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). 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 results while, for what concerns L10, the uncertainty is missing. Fig. 6 shows the corresponding cumulative distribution.

From these examples 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.

Concerning the PC3.5 connector, we have to declare an accidental breaking of the sensor. For this reason we can not consider these data reliable, so the analysis presented has been limited to the PCN 7 mm connectors data up to 18 GHz.

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. In conclusion the Microwave Power Measurement Comparison

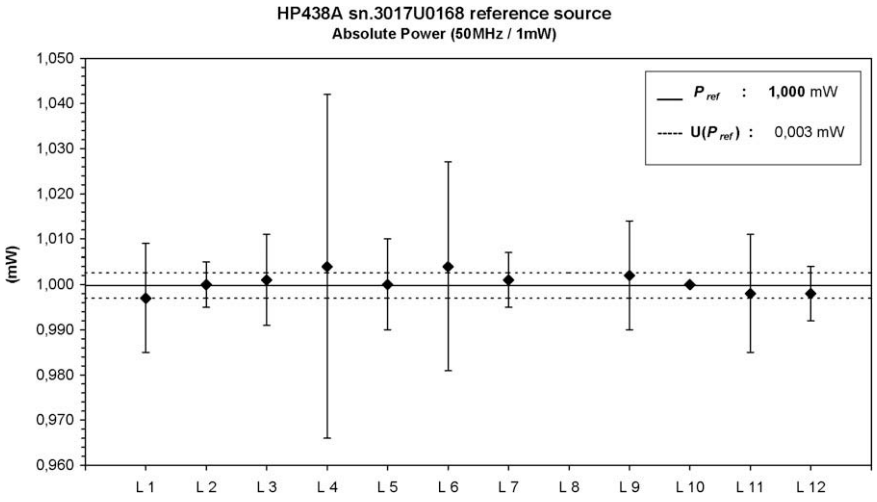

Fig. 5. Results for the absolute power measurement of the traveling standard reference source.

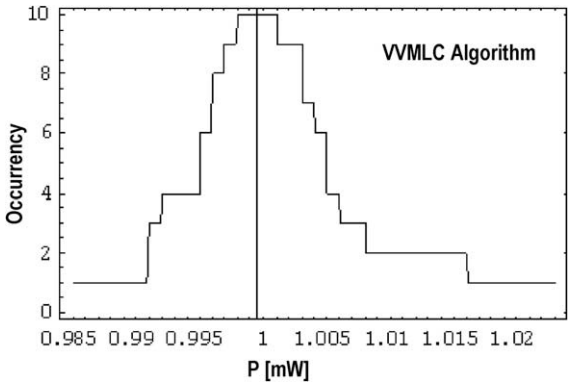

Fig. 6. Cumulative distribution for the power measurement of the traveling standard reference source.

exercise SIT.AF-01 has been a success, at least in the PCN 7 mm range of measurements, because 70% of the accredited laboratories are consistent with the uncertainties declared and with INRiM measurements, while only 30% of them presents light discrepancies only for some frequencies.

Acknowledgements

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, P. Terzi, Confronto nazionale interlaboratorio SIT.AF-01. Potenza AF in linea di trasmissione coassiale da 50 MHz a 26,5 GHz, SIT Technical 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, E. Vremera, A new microcalorimeter for measurements in 3.5 mm coaxial line, IEEE Trans. Instr. Meas. 52 (2) (2003) 320–323.
[4] H.S. Nielsen, Determining Consensus Value in interlaboratory comparison and proficiency testing, NCSLI Newsl. 44 (2) (2004).
[5] Guide to the expression of the uncertainty in measurement, ISO, Geneva, 1993.
[6] Mathematica. Available from: <http://www.wolfram.com/>.
[7] European Co-operation for Accreditation, EAL Interlaboratory Comparison (previously EAL-P7) Withdrawn, EA-2/03 rev.1.