Consistency plots: a simple graphical tool for investigating agreement in key comparisons

Stephen L. R. Ellison

Received: 27 May 2022 / Accepted: 26 July 2022 / Published online: 5 September 2022
© The Author(s) 2022

Abstract
A simple graphical display is described for investigating agreement among interlaboratory data with reported uncertainties. The plot consists of a measure of agreement—the significance level of a test for significant pairwise difference, adjusted for multiple comparisons—plotted as an image in which significance is represented by colour or intensity. This provides an easily interpretable graphical presentation in which the degree of consensus can be judged and in which anomalies are easily visible. The construction of the plot is discussed, with attention to the choice of adjustment for the size of the data set in checking for anomalies. The advantage over visual inspection using error bars is discussed, and some examples from metrology comparisons are presented.

Keywords Interlaboratory comparisons · Key comparisons · Statistics · Graphical methods

Introduction

Many inter-laboratory metrology comparisons, including many of the key comparisons required by the CIPM MRA [1, 2], generate data in the form of reported results with stated uncertainties. A critical early step in the assessment of such data is the identification of anomalies [3]. This is essential to help identify and resolve possible technical issues before any summary or performance statistics are calculated.

Graphical inspection is widely recommended for initial data inspection, not least because “there can be surprises in the data” [4, p. 8]. However, the most common plot of key comparison data—the familiar dot-and-bar plot—is not a reliable indication of mutual consistency or validity of uncertainty estimates. Other useful graphical methods have been suggested; for example, Duewer’s Concordance/Apparent Precision plot [5] and the Naji and (recently) Naji 2 plots [6–8]. These, however, rely on a reasonably reliable reference value. Many such studies do not have the advantage of an independently determined reference value with small uncertainty, leaving the choice of reference value open to debate and compromising graphical assessments that rely on a central location. This makes it important to seek alternative graphical tools that do not depend heavily on a reliable reference value.

Some procedures have been suggested that address the lack of a reliable reference value by reverting to pairwise comparisons (see, for example, [9, 10]), and some can provide graphical output. A plot of the median scaled difference, for example, can provide a useful indicator of location/uncertainty anomalies without any general location estimate [10]. It can, however, be useful to understand the structure and grouping of results in more detail, and for that purpose a more detailed presentation is useful.

In this paper, the shortcomings of the dot-and-bar plot, as an indicator of significant disagreement, are reviewed before offering an alternative graphical indicator of pairwise inconsistency that can be used for quick and reliable visual identification for important pairwise differences.

Methods

All statistical operations used the R statistical programming environment, version 4.1.3 [11]. Comparison plots used the metRology package for R, version 0.9-28-1 [12].
A critique of the dot-and-bar plot

The usual tool for displaying raw key comparison data is the dot-and-bar graph (sometimes—particularly when the quantity axis is horizontal—also called a ‘caterpillar plot’). An example of a typical dot-and-bar chart is shown in Fig. 1.

The plot consists of individual results and associated uncertainty intervals. A central value is often included in the plot, which may have its own uncertainty (usually shown with an additional pair of lines bracketing the central line). This type of plot neatly displays the reported data, but as a tool for identifying inconsistencies it falls short in several respects.

First, the plot is sometimes constructed using standard uncertainties for the error bars. Although an accurate visual presentation of standard uncertainties, this gives no useful visual indication of the significance of any differences.

Second, visual comparison of plotted uncertainty intervals is often ambiguous. While cases (a) and (b) in Fig. 2 can always be taken as indicating a lack of significant difference at the chosen confidence level, cases (c) and (d) are less clear. For just two observations, cases like (c) and (d) may indicate significance anywhere between about 84 % and 99.4 %. These levels arise, respectively, when both observations are just outside each other’s intervals, and when the intervals are just in contact (case (d) in Fig. 2), both with $u_1 = u_2$. Nor is it straightforward to recognise the degree of overlap required for a chosen level of significance, as that depends on the relative sizes of the two uncertainties. As Fig. 3 shows, even for the simple case of only two observations, exact 95 % significance occurs with quite different extents of overlap, depending on the two uncertainties involved.

Third, visual inspection of a dot-and-bar plot for differences implicitly seeks differences among all possible pairwise comparisons; it is, implicitly, conducting multiple simultaneous hypothesis tests. For a study with $n$ participants, there are $n(n-1)/2$ distinct pairwise comparisons. This is a well-known problem in general statistics. D’Errico [14] provides a useful review of the multiple comparison problem in the context of metrology; Elster and Toman [15] allude to it in the context of key comparison data evaluation.

![Fig. 1 Dot-and-bar plot of results for cadmium in a sediment, from key comparison CCQM-K13 [13]. Error bars show expanded uncertainty with coverage factor $k = 2$. The horizontal central line is the median for the data set. The arrow indicates the location of an extreme result for the leftmost laboratory, at 2.13 µmol kg$^{-1}$.](image1)

![Fig. 2 The figure shows four pairs of observations (⋆) with 95 % coverage intervals (error bars). (a) Both observations within both intervals (no significant difference at 95 %); (b) one observation inside the other observation’s interval (no significant difference at 95 %); (c) both points just outside one another’s intervals; (d) intervals just in contact, with zero overlap](image2)

![Fig. 3 The figure shows pairs of points $(x_1, x_2)$ (⋆) separated by exactly the critical difference for 95 % confidence given the uncertainties shown by the bars. Error bars show the 95 % confidence interval for each observation. Numerical values below each pair show the ratio $u_1/u_2$ of the uncertainties associated with each value in the pair. In all three cases, a two-sided $t$-test returns exactly $p = 0.05$](image3)
Briefly, multiple pairwise comparisons drastically increase the probability of type 1 error, usually known in this context as the family-wise type 1 error rate (FWER). Type 1 error is incorrect rejection of a null hypothesis, that is, a ‘false positive’. In the metrology comparison context, H0 corresponds to the hypothesis of no difference in (true) mean values. The FWER is the probability of incorrect rejection of this general null hypothesis, based on one or more apparently significant findings among a ‘family’ of multiple pairwise tests. The FWER for an uncorrected multiple pairwise test increases rapidly with the number of laboratories. In a ten-lab study where laboratories are performing identically, at least one apparently significant difference can be expected by chance if all pairwise tests are carried out at the usual 95% level. This leads to incorrect conclusions and wasted time following up false positives.

Over-interpretation because of implicit multiple comparisons, and under-interpretation because of the focus on interval overlap, can partly cancel in modest studies. But the actual false positive rate still depends on the size of the study. In very small studies, visual inspection of the dot-and-bar plot underestimates the significance of differences, and for larger studies (ten or more) it quickly starts to overstate significance owing to the multiple comparison problem.

Finally, the standard dot-and-bar plot may need careful and lengthy inspection to identify all the anomalous results. Even when a reliable independent reference value is available, overlap between individual expanded uncertainty intervals and the corresponding interval for the key comparison reference value (KCRV) suffers exactly the same interpretation ambiguities as shown in Figs. 2 and 3. Without such a reference value, the viewer must visually compare each laboratory result with all of the others in turn, taking account of the ambiguities above, to identify possible concerns.

The dot-and-bar plot, then, remains a generally useful tool for presenting metrology comparison data, but it is not easily interpreted for marginal cases. For simple pairwise comparisons, it encourages under-interpretation of differences; for detecting general inconsistency, it can under- or over-interpret, depending on the size of the study and the actual inspection criterion used.

**The pairwise consistency plot**

One solution to the problems above is the ‘consistency plot’ proposed here. The plot consists of the p values for all tests for significant pairwise difference, adjusted for multiple comparisons, plotted as an image in which the degree of disagreement is represented by colour or colour intensity. This provides a very quick and reliable visual indication of mutual inconsistencies among results with associated uncertainties.

By way of example, consider the data displayed in Fig. 1. The task is, first, to check whether the reported uncertainties explain the observed dispersion and, second, to determine which results might be responsible for any apparent overdispersion.

A commonly recommended check for whether key comparison results are mutually consistent within their uncertainties is a Chi-squared test, based on differences between some central estimate and each result, each divided by the uncertainty of the difference [16]. For the data in Fig. 1, even omitting the obvious low outlier at 2.13 μmol kg⁻¹, a Chi-squared test with the usual weighted mean as the central estimate returns a Chi-squared value of 167 for 12 degrees of freedom and a corresponding p value of 1.61 × 10⁻²⁹—convincing evidence of appreciable overdispersion. While this provides a convenient summary test, however, it tells us nothing about which results are causing any excess dispersion. Inspection of Fig. 1 suggests at least laboratories 13 and 14 (as well as laboratory 2), but others also seem inconsistent with the provisional location, if not necessarily with one another. Checking whether all are singly or jointly responsible for the high Chi-squared value is a slow process. Instead, this more detailed question can be addressed by a multiple pairwise test; that is, each result is tested against every other result for a significant difference, with adjustment for multiple comparisons.

For most purposes in metrology, a z- or t-test is often sufficient for comparing a pair of observations with uncertainty. For independent results x₁ and x₂ with uncertainties u₁ and u₂, respectively, the usual t- or z-tests calculate the test statistic

\[
\tau = \frac{|x_1 - x_2|}{\sqrt{u_1^2(x_1) + u_2^2(x_2)}}
\]  

(1)

The uncertainty term in the denominator can be adjusted for correlation if needed. The statistic τ can be interpreted either by comparison with an appropriate critical value or by inspecting a p value derived from the normal or t distribution, depending on whether degrees of freedom are important. With appropriate adjustment for the number of pairwise comparisons, this not only provides a test for inconsistency; the list of test statistics or, equivalently, their p values, shows which pairs of results appear to show significant difference.

The consistency plot proposed here is simply a plot of these p values, adjusted for multiple comparisons, with colour or colour intensity chosen for familiar thresholds for significance. The consistency plot for the data in Fig. 1 (data in table S1, supplementary information) is shown in Fig. 4. Interpretation is simple; a strong colour at the intersection for two laboratories (laboratories appear along both x- and y-axes) indicates a significant difference between the two, in this case after Holm adjustment.
It is of course possible to mark the plot to show these as an exception; here, no special indication is given, in order to retain the focus on discrepancies.

Third, in the plot shown here, laboratories have been shown in order of ascending measured value. This accounts for the typically empty central region, in which eight or nine laboratories show no significant pairwise difference in this study. This makes it relatively simple to identify what appears to be a mutually consistent subset of results. It can also suggest the presence of multiple such subsets. For example, the empty region at top right of Fig. 4 might suggest a small but mutually consistent subset of three laboratories, though in this case it would be prudent to note the relatively large differences and uncertainties among those results. A different ordering could, however, change the plot appearance considerably and might offer advantages in some cases; the effect of ordering is discussed below, in connection with the example from CCQM-K86c.

Finally, with a Holm adjustment, the appearance of any significant difference in the plot is cause to reject a null hypothesis of equal means and accurate uncertainties. This follows because a Holm adjustment strongly preserves the FWER [17]; at the null, sampled data sets will generate one or more significant Holm-corrected p values with essentially the same frequency as a Chi-squared test will show significant inconsistency. The method of p value adjustment is, however, an important choice and is discussed further below.

**Choice of p value adjustment for multiple hypothesis tests**

Multiple test p value adjustment procedures all, to some extent, increase the individual pairwise test p values by an amount dependent on the intended outcome [18]. Adjustment methods can have different objectives. The most common is to preserve the FWER. For example, the simple and well-known Bonferroni correction for multiple pairwise hypothesis tests [19], implemented as a p value adjustment, simply requires multiplication of all p values by the number of pairwise tests and capping the returned p values at 1.0. To a good approximation this maintains the FWER of (say) a Chi-squared test on the complete data set. However, it will often do so at the expense of missing some marginally significant pairwise differences that might individually be worth closer inspection. As the data set size increases, this can quickly become problematic for a study organiser wishing to quickly identify multiple potentially anomalous points for closer inspection. Choice of p value adjustment method is therefore important.

Figure 5 illustrates the effect of several well-known adjustment methods. Two of the procedures shown, the Bonferroni and Holm adjustments, strongly preserve FWER. Of the two, the Bonferroni adjustment (Fig. 5a) is simplest but
least powerful for identifying real pairwise differences. As the figure shows, the Bonferroni adjustment generates the smallest number of apparently important pairwise differences. Holm’s method (Fig. 5b) is more powerful for identifying important pairwise differences, but still strongly preserves FWER; it is consequently to be recommended over Bonferroni’s method for most purposes.

An alternative objective for adjustment, however, is to balance high power for detection of real pairwise differences with good control of the FWER at the null. A widely used procedure that achieves this is that of Benjamini and Hochberg [20]. Instead of FWER, this procedure controls the false discovery rate; that is, the probability that an apparently ‘significant’ difference is just chance. At the null it provides good FWER control, but it also correctly identifies a higher proportion of real differences. For real differences it has consistently higher power than Holm’s procedure [20]. As Fig. 5c) shows, this generates more potentially important differences among the cadmium data results. Where the principal intention is to identify possible differences for closer investigation, then, the Benjamini–Hochberg procedure should be preferred.

The figure includes a comparison with a plot using no $p$ value adjustment, for completeness (Fig. 5d). In this case, the change from the Benjamini–Hochberg adjustment is modest. This is a fair indication that most of the ‘significant’ differences in Fig. 5d) are genuine; recall that the Benjamini–Hochberg procedure controls the false discovery rate, so only about 5% of ‘significant’ results, after adjustment, will be incorrect. In other cases, where participants perform consistently, failing to adjust will lead to an unacceptably high proportion of apparently significant differences, and follow-up will waste study coordinator and participant time.

**Further examples**

The utility of these consistency plots is perhaps best illustrated by example. Here, two further examples are presented.
Figure 6a) shows reported results and uncertainties from a CCQM pilot study on electrical conductivity [21]; for the data, see table S2 in supplementary information. In this early study, the reported uncertainties vary considerably due to different measurement technologies in use at the time. As a result, the dot-and-bar plot in Fig. 6a) needs expansion (see inset) to inspect results for consistency and, even then, inspection remains challenging. The corresponding consistency plot in Fig. 6b), however, shows immediately that results from laboratories 2 through 6 differ significantly from those for laboratories 10 through 13, with laboratories 7 and 8 inconsistent with 11 and 13, and that laboratory 2 shows additional significant differences (adjusted $p$ value $< 0.05$) compared to laboratories 5, 6, and 8. The overall picture is of agreement among results from laboratories 1, 3 through 9, and 14, but appreciable differences between the central group (3 through 9) and other results.

Figure 7a) shows results and uncertainties for one of two materials examined in key comparison CCQM-K86c, a comparison on determination of a specified gene target ratio in *Brassica napus* (canola) [22]; for the data, see table S3 in supplementary information. The CCQM-K86 range of studies [22–24] supports the quantification of proportion of genetically modified (GMO) material present in otherwise wild-type materials. For this material, the nucleic acid target was a fragment introduced in modification event RT73,
random ordering in a consistency plot generates easily inter-
mentary material.
Fig. 7a), in ascending value order, can be found in supple-
by ascending or descending reported value. A variation of
ful to start with a plot in which laboratories are ordered
pairwise differences. For that reason, it will often be use-
obscure important patterns when there are many significant
pretable conclusions when there are relatively few signifi-
cation of the two, and possibly three, distinct groups which
essentially random ordering in Fig. 7a) gives no strong indi-
the canola genome. The consistency
plot in Fig. 7b) identifies the resulting disagreement clearly,
via the high proportion of low \( p \) values at top right and lower
left; there is essentially no agreement across the two groups.
The plot also confirms that the two lowest values disagree
consistently with essentially all other participants, despite
appreciable overlap in confidence intervals for laboratories
K08 and K10.

The CCQM-K86c example also illustrates the importance
of an appropriate choice of data ordering in such plots. The
essentially random ordering in Fig. 7a) gives no strong indi-
cation of the two, and possibly three, distinct groups which
are visible in Fig. 7b). It is also worth noting that, while
random ordering in a consistency plot generates easily inter-
pretable conclusions when there are relatively few significa-
cant pairwise differences, random or arbitrary ordering can
obscure important patterns when there are many significant
pairwise differences. For that reason, it will often be use-
ful to start with a plot in which laboratories are ordered
by ascending or descending reported value. A variation of
Fig. 7a), in ascending value order, can be found in supple-
mentary material.

Limitations and alternatives

The pairwise consistency plot described here is much bet-
ter than the dot-and-bar plot for identifying significant
differences among results with uncertainties. In the form
suggested here, however, it has two limitations that may
sometimes be important. First, it might not scale well to
much larger data sets, such as a large proficiency testing
exercise with hundreds of results. Although it will con-
tinue to identify inconsistency, it will become difficult to
identify the specific participants responsible. Secondly,
although the principle of displaying adjusted \( p \) values
is relatively general, generating these becomes intricate
when individual reported uncertainties are associated
with appreciably different distributions. For small relative
uncertainties, this is rarely a serious issue, but for rela-
tive uncertainties much over 20 %, asymmetry can become
important and it may not be straightforward to provide
pairwise hypothesis tests for difference in location.

Scaling is rarely an issue for metrology comparisons
in the CIPM framework, as most comparisons are rela-
tively small. It is also noteworthy that the dot-and-bar plot
itself becomes crowded and hard to interpret as study size
increases. For much larger interlaboratory comparisons,
other tools are likely to be needed.

The problem of differing uncertainty distributions is
more difficult to address. In principle, the probability of
a given observed difference in location between observa-
tions from almost any pair of distributions can be obtained
via simulation methods. In practice, this can be intracta-
ble; for example, the variance of many non-normal dis-
tributions depends on a true location, and constructing a
null hypothesis for a pairwise test can be difficult. Meas-
ures other than hypothesis test probabilities might then be
more appropriate. For example, although not examined
here, plots based on indicators such as Kullback-Leibler
divergence [26] and related measures could prove informa-
tive. In addition, it should not be forgotten that Cofino
and others have long used a distribution overlap integral
as a basis for a robust estimate of location [27, 28], and
plots of the pairwise overlap matrix are very similar in
appearance to the plots described here [29]. The principal
difference is that the consistency plots in the present paper
provide a direct indication of pairwise test significance,
whereas overlap integrals have not so far been used for
that purpose.

Conclusions

As shown above, a consistency plot of the kind suggested
here is a useful addition to the range of graphical methods
for inspecting key comparison and similar data. With the
right choice of \( p \) value adjustment, the plot provides both a
test for general inconsistency and a clear indication of the
particular differences that may need to be examined more
closely. It has the additional advantage of not depending
on a choice of estimator for a central value, making it
particularly useful in initial review of studies with no inde-
pendent reference value.

An implementation of the plot is available in the
metRology [12] package for R. A simple spreadsheet
implementation, using a Bonferroni adjustment, is avail-
able as electronic supplementary material for MS Excel
versions that support the MS Excel SORTBY function
(MS 365, available from mid-2020).

Supplementary Information The online version contains supplemen-
tary material available at https://doi.org/10.1007/s00769-022-01520-z.
Acknowledgement  Publication of this paper was funded in part by the UK government Department for Business, Energy & Industrial Strategy (BEIS).

Open Access  This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

1. CIPM (2003) Mutual recognition of national measurement standards and of calibration and measurement certificates issued by national metrology institutes. Bureau International des Poids et Mesures, https://www.bipm.org/en/cipm-mra/cipm-mra-documents
2. CIPM (2021) CIPM MRA-G-11: measurement comparisons in the CIPM MRA – Guidelines for organizing, participating and reporting. Guidance document, Comité international des poids et mesures, https://www.bipm.org/en/cipm-mra/cipm-mra-documents, (accessed 2022-07-21)
3. CCQM (2013) Estimation of a consensus KCRV and associated degrees of equivalence, version 10. CCQM Guidance note, Comité Consultatif pour la Quantité de Matière, https://www.bipm.org/cc/CCQM/Allowed/19CCQM13-22_Consensus_KCRV_v10.pdf
4. Cleveland WS (1994) The elements of graphing data. AT &T Bell Laboratories, Murray Hill, N.J, rev. ed ed., ISBN 9780963488411
5. Duewer DL, Kline MC, Sharpless KE, Thomas JB, Gary KT, Sowell AL (1999) Anal Chem 71:1870–1878. https://doi.org/10.1021/ac981074k
6. Robouch P, Younes N, Vermaercke P (2003) In Richter D, Wöger W, Hässelbarth W (eds.), Data analysis of key comparisons, pp 149–160, Physikalisch-Technische Bundesanstalt, Berlin, Germany, ISBN 3-89701-933-3, ISSN 0942-1785
7. Aregbe Y, Harper C, Nørgaard J, Smet MD, Smeyers P, Van Nevel L, Taylor PDP (2004) Accred Qual Assur 9:323–332. https://doi.org/10.1007/s00769-003-0693-x
8. Cordeiro F, Emons H, Robouch P (2022) Accred Qual Assur 27:145–153. https://doi.org/10.1007/s00769-022-01496-w
9. Douglas RJ, Steele AG (2006) Metrologia 43:89–97. https://doi.org/10.1088/0026-1394/43/1/013
10. Ellison SLR (2018) Metrologia 55:840–854. https://doi.org/10.1088/1681-7575/aae610
11. R Core Team (2022) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, https://www.R-project.org/
12. Ellison SLR (2018) metRology: support for metrological applications. https://r-forge.r-project.org/projects/metrology/. R package version 0.9-28-1
13. Papadakis I, Taylor PDP, Aregbe Y (2002). Metrologia. https://doi.org/10.1088/0026-1394/39/1a/22
14. D’Errico GE (2013) J Phys: Conf Series 459(012):041. https://doi.org/10.1088/1742-6596/459/1/012041
15. Elster C, Toman B (2013) Metrologia 50:549–555. https://doi.org/10.1088/0026-1394/50/5/549
16. Cox MG (2002) Metrologia 39:589–595. https://doi.org/10.1088/0026-1394/39/6/10
17. Holm S (1979) Scand J Stat 6:65–70
18. Wright SP (1992) Biometrics 48:1005–1013. https://doi.org/10.1088/0026-1394/39/6/10
19. Bonferroni CE (1935) Il calcolo delle assicurazioni su gruppi di teste. Studi in Onore del Professore Salvatore Ortu Carboni
20. Benjamini Y, Hochberg Y (1995) J Royal Stat Soc Ser B 57:289–300
21. Jensen HD, Dam N-E (2001) Report on CCQM-P22. Tech. rep, Danish Fundamental Metrology Ltd, Lyngby, Denmark
22. Mester Z, Corbisier P, Ellison SLR, Gao Y, Niu C, Tang V, Lee FW, Pérez-Uruquiza M, Suírez AR, Burns M, Milavec M, Wiangnon K, Griffiths KR, McLaughlin JLH, Shibayama S, Takatsu A, Akgoz M, Vonsky M, Runov A, Guerrero JEL (2020) Metrologia 57(08):004. https://doi.org/10.1088/0026-1394/57/1a/08004
23. ...Corbisier P, Vincent S, Schimmel H, Kortekaas AM, Trampmann S, Burns M, Bushell C, Akgoz M, Akyurek S, Dong L, Fu B, Zhang L, Wang J, Uruquiza MP, Bautista JL, Garibay A, Fuller B, Baoutina A, Partis L, Emslie K, Holden M, Chum WY, Kim HH, Phunbua N, Milavec M, Zel J, Vonsky M, Konopelko LA, Lau TLT, Yang B, Hui MHK, Yu ACH, Viroonudompohl D, Prawetongson C, Wiangnon K, Takabatake R, Kitta K, Kawaharasaki M, Parkes H (2012). Metrologia. https://doi.org/10.1088/0026-1394/49/1a/08002
24. Dong L, Sue Z, Wang J, Tang VHM, Chum WWY, Lee F, Sin DWM, Pérez-Uruquiza M, Burns M, Ellison SLR, Parkes H, Milavec M, Prawetongson C, Griffiths KR, McLaughlin JLH, Shibayama S, Akyurek S, Akgoz M (2018) Metrologia 55(08):017. https://doi.org/10.1088/0026-1394/55/1a/08017
25. Demekete T, Eng M (2018) Biomol Detect Quantif 15:24–29. https://doi.org/10.1016/j.bdq.2018.03.002
26. Kullback S, Leibler RA (1951) The Ann Math Stat 22:79–86
27. Cofino WP, van Stokkum IH, van Steenwijk J, Wells DE (2005) Appl Spectrosc 59:2304–2311. https://doi.org/10.1366/0003702053649739
28. Molenaar J, Cofino WP, Torfs PJ (2018) Chemomet Intell Lab Syst 175:65–73. https://doi.org/10.1016/j.chemolab.2018.01.003
29. Cofino WP, van Stokkum IH, van Steenwijk J, Wells DE (2005) Anal Chimica Acta 533:31–39. https://doi.org/10.1016/j.aca.2004.11.008

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.