RADIOCARBON, Vol 45, Nr 2, 2003, p 213–225

SECTION 5: MEASURES OF PRECISION AND REPRODUCIBILITY

5.1 DUPLICATES

The design of FIRI included 3 pairs of duplicate samples: A and B (Kauri wood) near background, D and F (Belfast wood) around 50 pMC, and G and J (barley mash) at 111 pMC. Why include duplicates? Duplicates by their nature allow us to explore the within-lab variability and to assess whether the quoted errors are representative. We can also explore the differences as a function of the sample activity. In this section, we explore the differences between the duplicates. We also consider some different graphical presentations. First, we summarize the differences, then graphically explore the boxplot (to consider the distribution of differences), then a scatterplot of the duplicate pair (to show correlation and reproducibility), and finally, a measure of agreement plot (Bland and Altman 1999). The horizontal axis in this final plot is the mean of the duplicate pair and the vertical axis is the difference in the duplicate pair. Agreement between the pairs would result in the points being randomly scattered around the horizontal zero line.

5.1.1 Summary Statistics for Duplicate Pairs

The summary statistics for the duplicates are shown below.

Table 5.1 Descriptive statistics: differences between duplicates (note: DF in yr BP)

Sample pair	N	Mean	Median	StDev	Min	Max
AB	54	0.0295	0.0000	0.2145	–0.66	0.531
GJ	71	–0.094	–0.080	1.085	–4.37	2.76
DF	79	17.4	17.0	97.3	–239	310

On average, the differences are close to zero, although it can be seen from the minimum and maximum that there is a wide scatter for sample pair GJ. For GJ, the largest difference between a pair of duplicates is just over 4 pMC, and for sample pair DF, the largest difference is 310 yr, both of which are small given the absolute activity/age of the sample. For Sample AB, the largest difference is 0.7 pMC, which is large given the near background activity for this sample. Each sample is now considered in more detail. The same pattern of analysis is repeated for the summaries by laboratory type (Tables 5.2–5.4). It is worth noting that 2 out of the 3 largest differences for the duplicates are reported by LSC laboratories.

Table 5.2 Descriptive statistics: AB differences by laboratory type

Lab type	N	Mean	Median	StDev	Min	Max
AMS	21	0.0436	0.0000	0.1234	–0.2	0.36
GPC	14	0.0662	0.0180	0.1621	–0.2	0.45
LSC	19	–0.0131	–0.0200	0.3105	–0.7	0.53

Table 5.3 Descriptive statistics: DF differences by laboratory type

Lab type	N	Mean	Median	StDev	Min	Max
AMS	25	8.7	17	68.9	–210	142
GPC	18	–2.7	5.0	96.4	–159	220
LSC	36	33.4	27.0	113.2	–239	310

https://doi.org/10.1017/S0033822200032616 Published online by Cambridge University Press

Table 5.4 Descriptive statistics: GJ differences by laboratory type

Lab type	N	Mean	Median	StDev	Min	Max
AMS	25	–0.2354	–0.1000	0.47	–1.1	0.8
GPC	17	–0.104	–0.080	1.31	–4.4	1.85
LSC	29	0.034	0.110	1.32	–3.0	2.8

5.2 SAMPLES A AND B

Figure 5.1 shows that the duplicate pair differences are, on average, zero. The scatterplot and agreement plots (Figures 5.2 and 5.3) both show that the points are quite widely scattered about the line of equality and the zero line, respectively, and that the scatter of the points increases with an increasing average pMC.

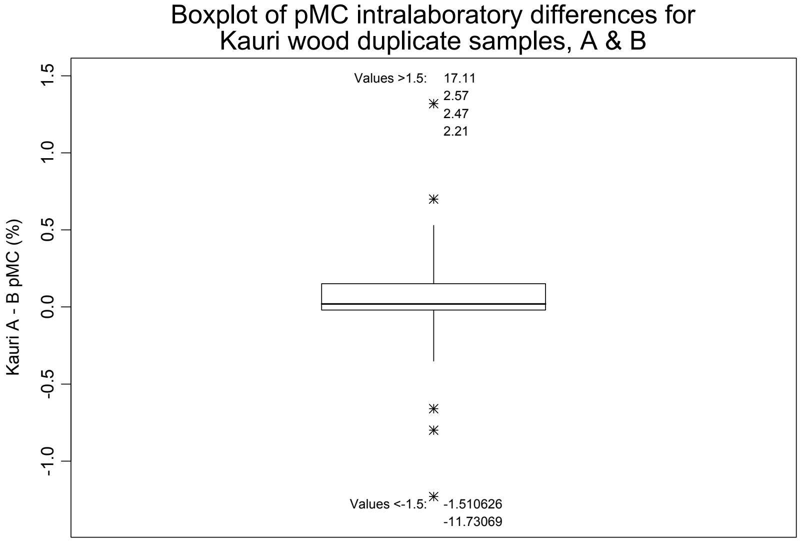

Figure 5.1 Distribution of differences (only differences <1.5 shown, uncensored results only)

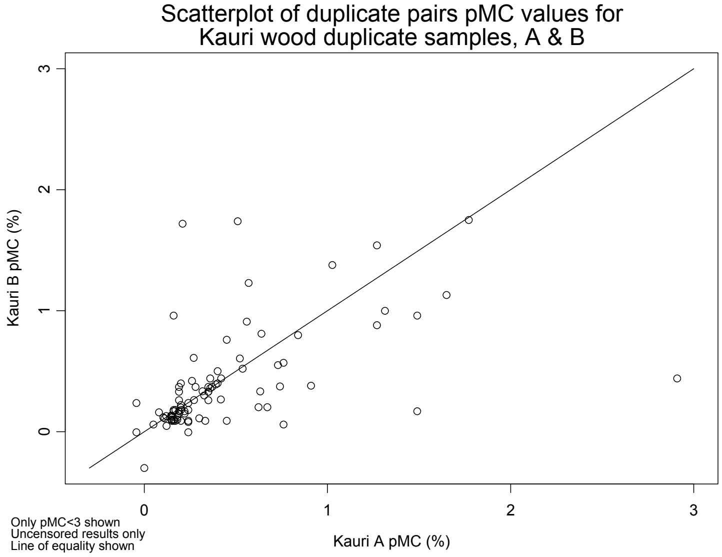

Figure 5.2 Scatterplot of duplicate pairs

https://doi.org/10.1017/S0033822200032616 Published online by Cambridge University Press

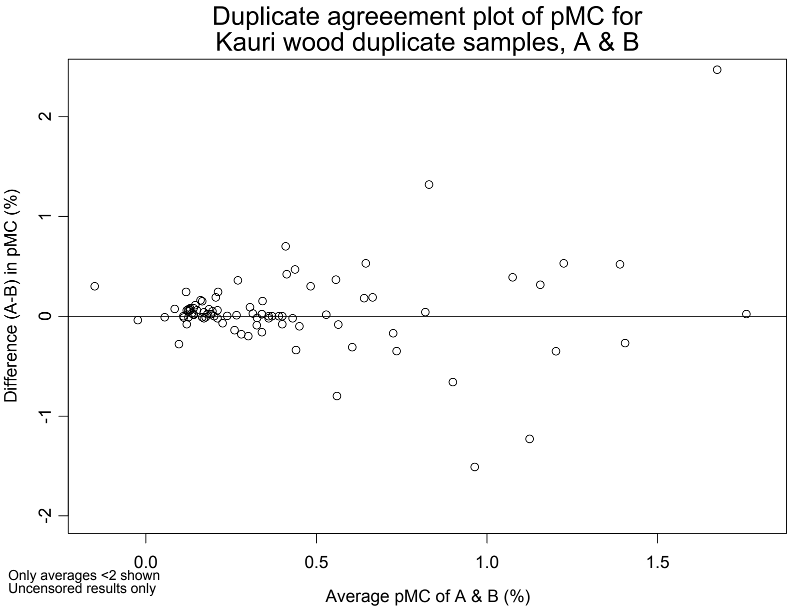

Figure 5.3 Agreement plot between duplicate pairs

5.3 SAMPLES D AND F

Figure 5.4 shows that the duplicate pair differences are, on average, zero. The scatterplot (Figure 5.5) shows that the pairs are quite widely scattered about the line of equality. Figure 5.6 shows a wide scatter around the zero line, suggesting that the difference is a function of the estimated age.

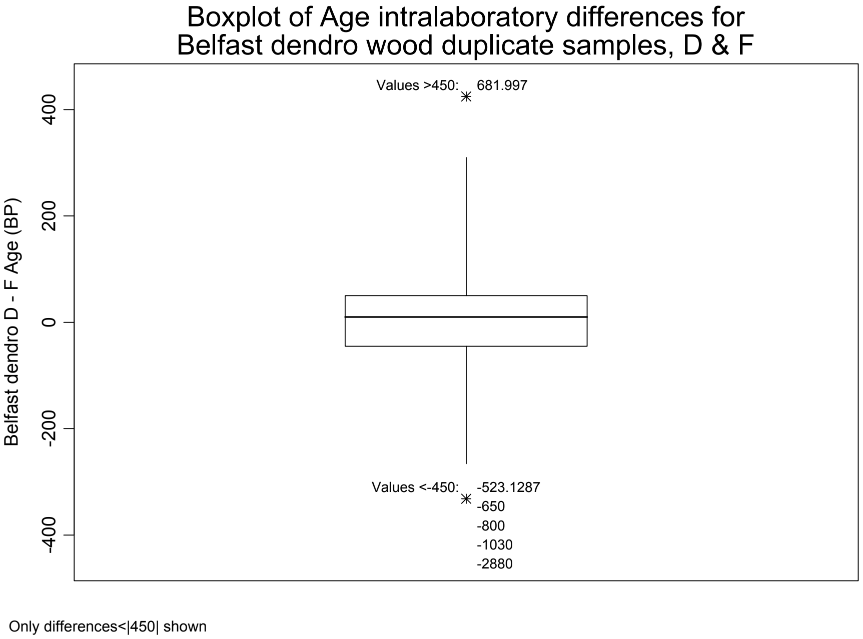

Figure 5.4 Distribution of differences

https://doi.org/10.1017/S0033822200032616 Published online by Cambridge University Press

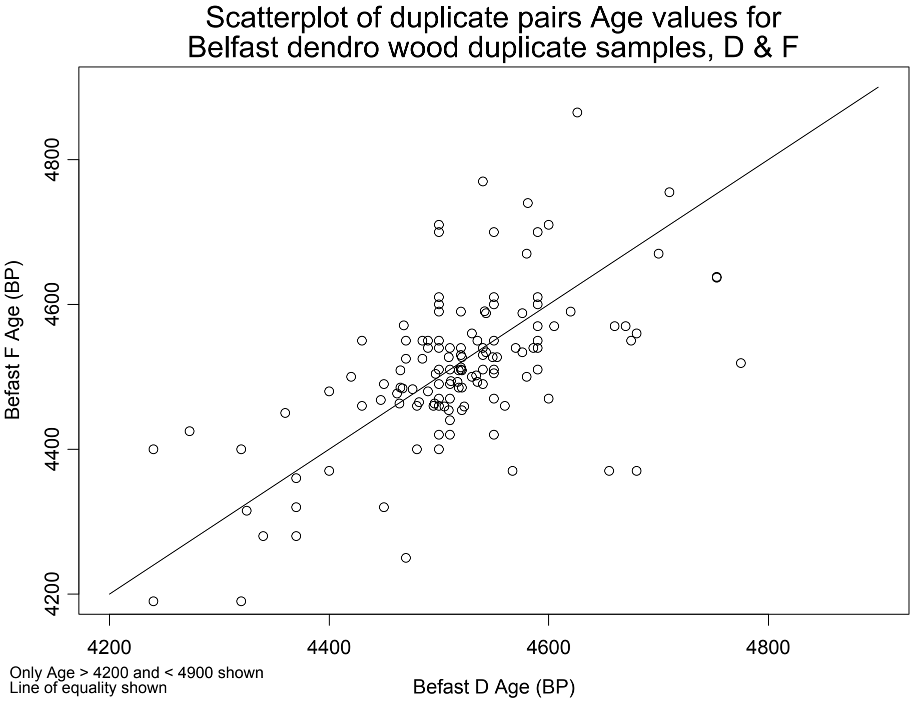

Figure 5.5 Scatterplot of duplicate pairs

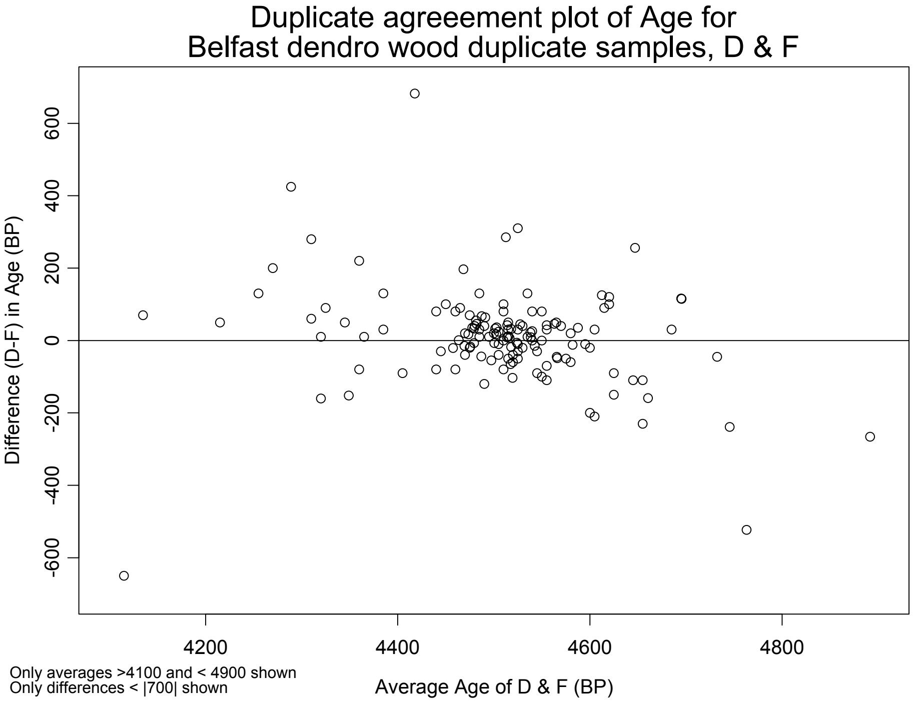

Figure 5.6 Agreement plot for duplicate pairs

https://doi.org/10.1017/S0033822200032616 Published online by Cambridge University Press

5.4 SAMPLES G AND J

Figure 5.7 shows that the duplicate pair differences are on average zero. The scatterplot (Figure 5.8) shows that pairs are quite widely scattered about the line of equality. Figure 5.9 shows a wide scatter around the zero line, with a number of outliers.

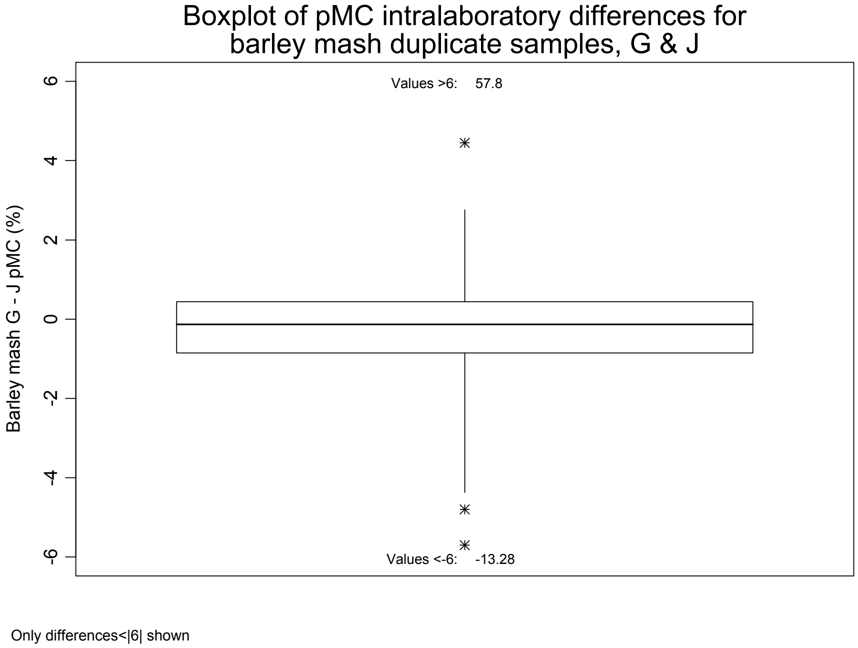

Figure 5.7 Distribution of differences

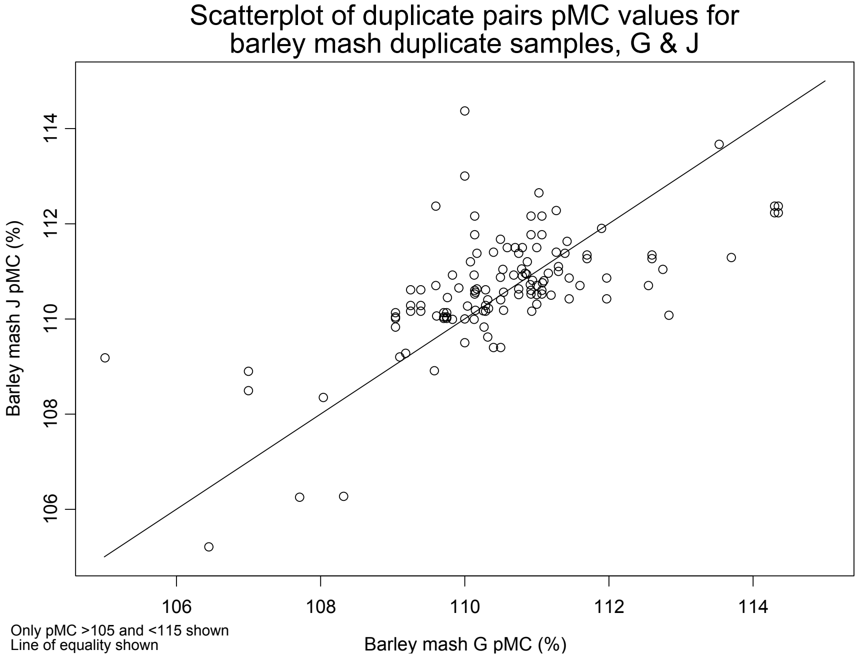

Figure 5.8 Scatterplot of duplicate pairs

https://doi.org/10.1017/S0033822200032616 Published online by Cambridge University Press

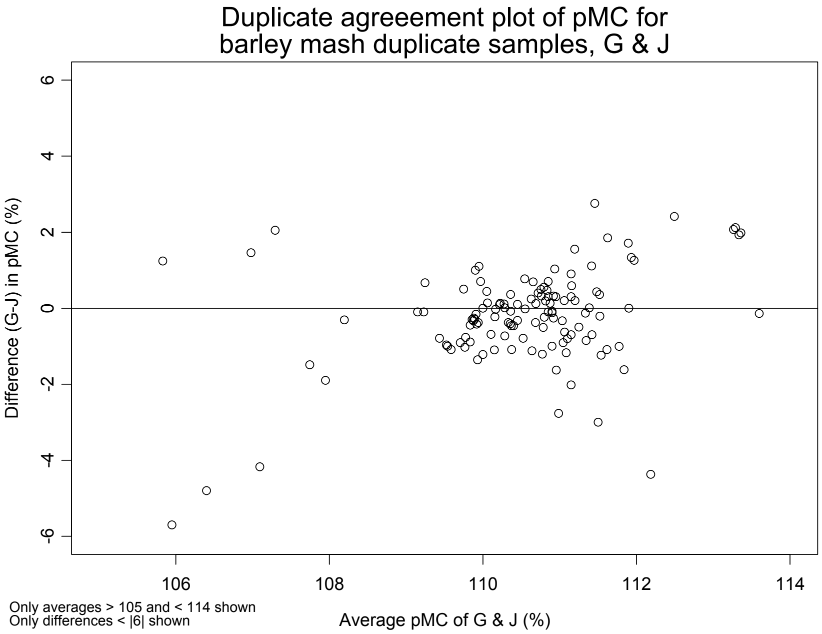

Figure 5.9 Agreement plot between duplicate pairs

5.5 QUOTED ERRORS

In addition, the duplicate results can also be used to assess the validity of the quoted errors. For each duplicate pair, the square of the difference, divided by the estimated standard deviation of the difference (deviance), should have a specific statistical distribution and name the Chi-squared distribution with 1 degree of freedom (or parameter) if the quoted errors adequately describe the uncertainty in measurement and, hence, the scatter in the differences. This theoretical distribution has a mean of 1 and a variance of 2 (standard deviation 1.4).

The tables below summarize the mean and standard deviation of the deviance each duplicate pair.

Table 5.5 Mean and standard deviation of the deviance for each duplicate pair

Sample pair	Mean	Standard deviation
AB	2.514	5.57
GJ	1.645	4.05
DF	2.220	4.76

Table 5.6a Mean and standard deviation of the deviance for duplicate pair AB by laboratory type

Sample pair AB	Mean	Median
AMS	3.36	1.19
GPC	1.54	0.185
LSC	2.220	0.27

Table 5.6b Mean and standard deviation of the deviance for duplicate pair GJ by laboratory type

Sample pair GJ	Mean	Median
AMS	0.85	0.19
GPC	2.28	0.45
LSC	2.05	0.41

https://doi.org/10.1017/S0033822200032616 Published online by Cambridge University Press

Table 5.6c Mean and standard deviation of the deviance for duplicate pair DF by laboratory type

Sample pair DF	Mean	Median
AMS	2.86	0.36
GPC	1.76	0.81
LSC	2.00	0.65

5.5.1 Comments

In conclusion, these tables show clearly that the distribution of the differences between each of the duplicate pairs does not correspond to the claimed uncertainties in the measurements, since the means and standard deviations do not agree with the theoretical values. This would suggest, in general, that the differences between the duplicates are more varied than would be expected, given the quoted errors.

5.6 REPRODUCIBILITY RESULTS

5.6.1 Repeatability and Reproducibility

Analyses performed on presumed homogeneous material do not yield identical results due to unavoidable random factors inherent in every measurement method. The repeatability and reproducibility of a standard measurement method are sufficient to describe the variability in a measurement method and can be estimated from an interlaboratory test. *Precision* is considered to be the closeness of agreement between independent measurements. *Repeatability* (r) refers to measurements made under identical conditions in one laboratory, while *reproducibility* (R) refers to measurements made in different laboratories, under different conditions. Reproducibility is the closeness of agreement between test results under conditions where the same method is used in different laboratories. The reproducibility quantifies the maximum variability in results. The samples used for such experiments should thus be sub-samples taken from 1 bulk sample, as is the case with the FIRI samples. In this section, we consider the following cases: a) the method is ^{14}C dating regardless of technique, and b) where we consider LSC, GPC, and AMS as 3 different methods.

We evaluate the repeatability and reproducibility values for a) the 3 pairs of duplicates (A, B; G, J; and D, F) and b) for all samples, but in this latter case, we need to modify the calculation method since we do not have replicate results, thus, we use the quoted errors.

The *reproducibility* value (R) is the value below which the absolute difference between 2 single results obtained under reproducibility conditions may be expected to lie with a probability of 0.95. A difference larger than *R* cannot be ascribed to random fluctuations and would warrant investigation of possible sources of systematic differences.

The method used is based on BS 5497 (1), however, outliers were defined by the 1.5 IQR method and removed before the BS 5497 (1) analysis was carried out. All results were converted to pMC to unify the interpretation.

5.6.2 Statistical Models

The basic model and estimating equations for *r* and *R* are given below:

$$\text{Model: } \mathrm{Y} = \mathrm{m} + \mathrm{B} + \mathrm{e}$$

where *Y* is the ^{14}C measurement, *m* is the general average for the particular material, *B* is the between-laboratory variation, and *e* is the random error.

https://doi.org/10.1017/S0033822200032616 Published online by Cambridge University Press

- *B* is assumed random in a reproducibility test and var(B) = σ^2_L
- *e* is also assumed random and within a single laboratory var(e) = σ^2_W
- We assume that σ^2_W is constant for all laboratories, with the average value σ^2_r
- The repeatability value *r* is 2.8 σ_r
- The reproducibility value *R* is 2.8 σ_R, where $\sigma_{R} = \sqrt{(\sigma^2_L + \sigma^2_W)}$

Estimation of *r* and *R* can be achieved from an intercomparison such as FIRI, where each sample can be considered as having one of *q* different levels of ^{14}C activity. The samples were sent to *p* different laboratories, which performed *n* analyses on each sample. In the case of FIRI for most samples, *n* is taken to be 1.

In the analysis for each sample separately, estimates of σ_r, σ^2_L and σ^2_R were calculated before evaluating *r* and *R*.

5.6.2 Analysis of the Duplicate Samples

The overall mean activity (m), the reproducibility measure (R), and repeatability measure (r) are shown for each material in Table 5.7.

Table 5.7 Repeatability and reproducibility

	AB	DF	GJ
m	0.348	56.991	110.603
R	0.749	1.551	2.613
r	0.451	1.047	1.728

The plots (Figure 5.10) below show the mean activity and standard deviation for the 3 pairs of duplicate samples. They show no obvious pattern between the mean and the standard deviation, but some extreme values are apparent (although they are not identified as outliers).

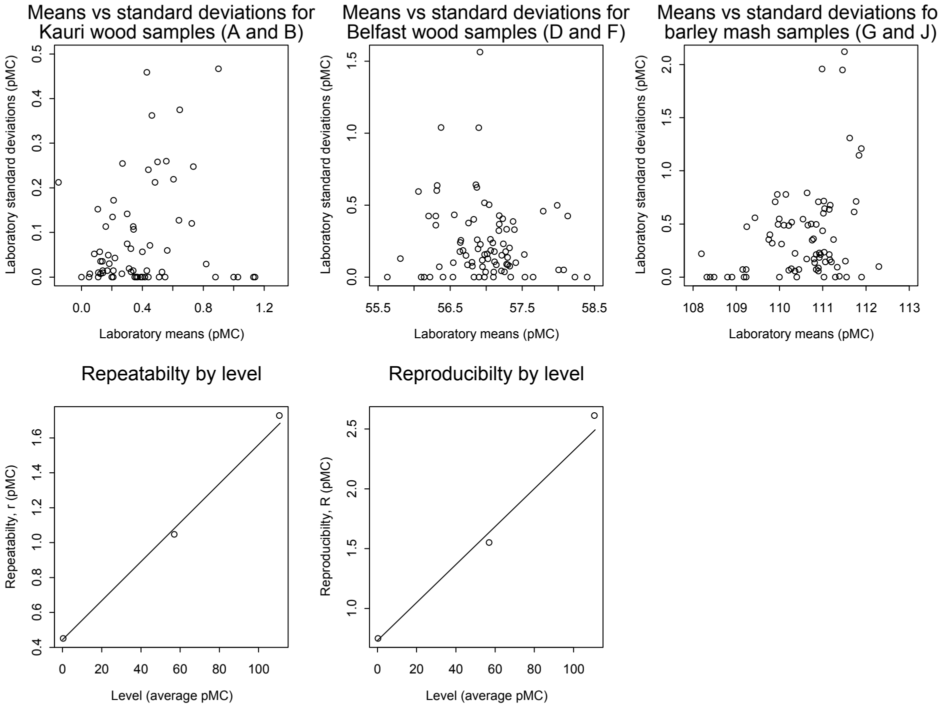

Figure 5.10 Scatterplots for duplicate samples

The last 2 plots show the strong linear relationship between *r* and *R* and the activity level.

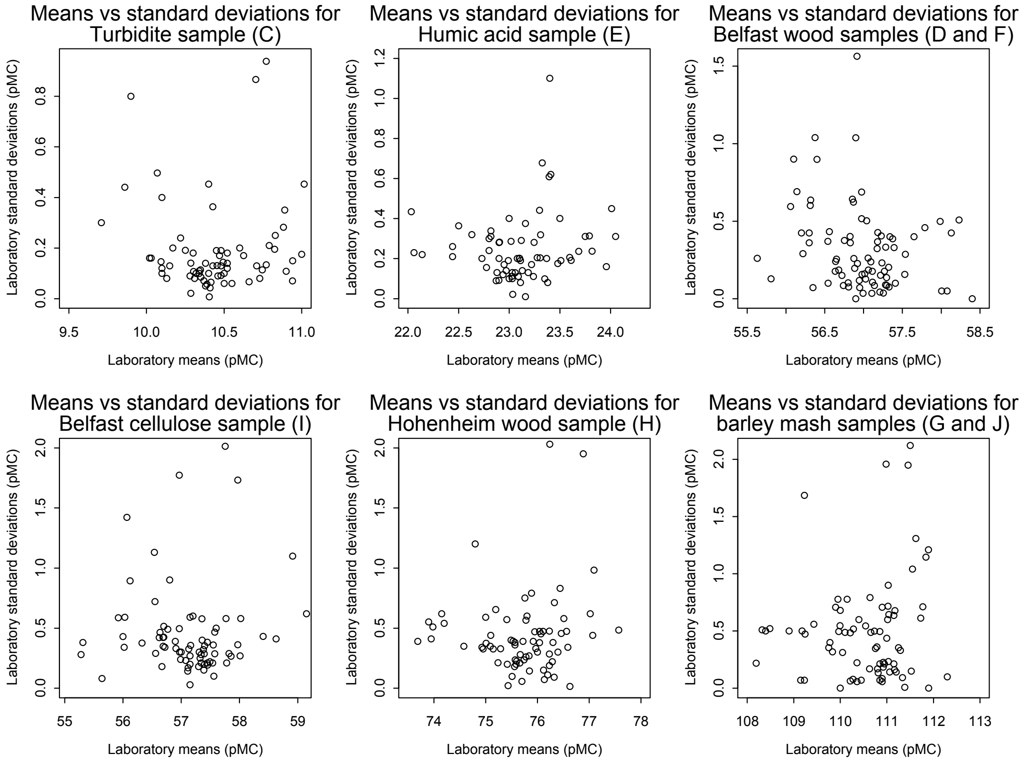

Figure 5.11 Scatterplot of means and standard deviations for all samples

The *R* values can be interpreted as the expectation that for any 2 randomly chosen measurements (i.e., laboratories), the absolute difference in their results should be less than 1.55 pMC (for a sample with an activity of 57 pMC), increasing to 2.6 for a sample with an activity of 110 pMC.

A similar analysis can be performed using all the samples (not simply the duplicate samples), however, here we need to modify the procedure such that the standard deviation previously calculated now must be estimated using the laboratory's quoted error for that sample.

5.6.3 C–J Results with Quoted Errors Used When No Replication Done

The quoted error is used as a substitute for the estimated standard deviation since we have no replicates.

Overall means (m), reproducibility measures (R), and repeatability measures (r) are given in Table 5.8.

Table 5.8 Reproducibility and repeatability for all samples

	C	E	DF	I	H	GJ
m	10.44	23.11	56.99	57.17	75.76	110.61
R	0.79	1.15	1.60	2.05	2.18	2.64
r	0.73	0.84	1.17	1.64	1.54	1.86

We can see quite clearly the dependence of *R* on the sample activity.

https://doi.org/10.1017/S0033822200032616 Published online by Cambridge University Press

5.6.4 Reproducibility for the Different Techniques

In this section, a similar analysis was performed, but for the laboratory types separately. Outliers, as defined by the 1.5 IQR method, are removed and all units are pMC.

5.6.4.1 Duplicate Results

Overall means (m), reproducibility measures (R), and repeatability measures (r) for the 3 measurements techniques are given in Table 5.9.

Table 5.9a AMS repeatability and reproducibility

	AB	DF	GJ
m	0.23	56.88	110.46
R	0.41	1.14	1.84
r	0.37	0.86	1.34

Table 5.9b GPC repeatability and reproducibility

	AB	DF	GJ
m	0.28	57.09	110.78
R	0.74	1.45	2.68
r	0.47	0.88	1.48

Table 5.9c LSC repeatability and reproducibility

	AB	DF	GJ
m	0.50	57.06	110.66
R	0.83	1.91	3.22
r	0.55	1.35	2.30

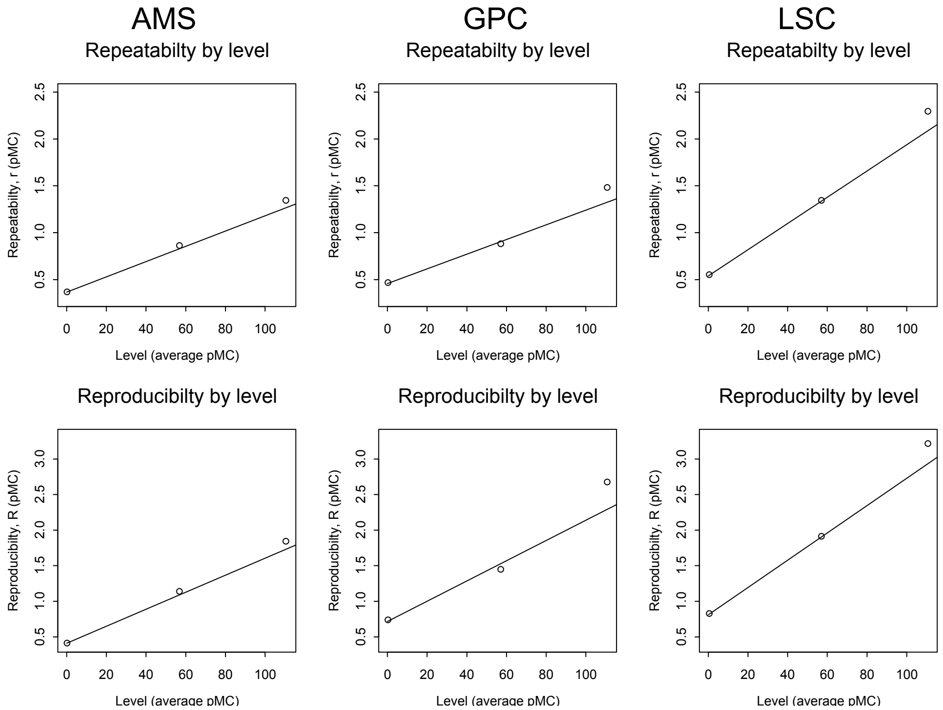

Figure 5.12 Repeatability and reproducibility for laboratory types

5.6.4.2 Comments

Large differences between techniques are observed, with AMS laboratories having lower reproducibility values compared to radiometric methods. LSC laboratories have higher repeatability values than the other techniques. Thus, for LSC, bigger differences in the results can be expected and we can expect more variation in the LSC results compared to AMS or GPC results.

5.6.5 C–J Results with Quoted Errors Used When No Replication for the Different Laboratory Types

Table 5.10a AMS repeatability and reproducibility

	C	E	DF	I	H	GJ
m	10.41	22.98	56.88	57.12	75.77	110.46
R	0.48	0.50	1.17	1.44	1.33	1.76
r	0.32	0.57	0.92	0.85	1.01	1.16

Table 5.10b GPC repeatability and reproducibility

	C	E	DF	I	H	GJ
m	10.40	23.24	57.09	57.53	75.82	110.78
R	0.90	1.28	1.47	1.28	2.64	2.68
r	0.54	1.04	0.94	1.05	1.15	1.48

Table 5.10c LSC repeatability and reproducibility

	C	E	DF	I	H	GJ
m	10.49	23.16	57.06	57.05	75.71	110.68
R	0.90	1.42	1.95	2.56	2.57	3.20
r	0.97	0.90	1.40	2.21	2.01	2.39

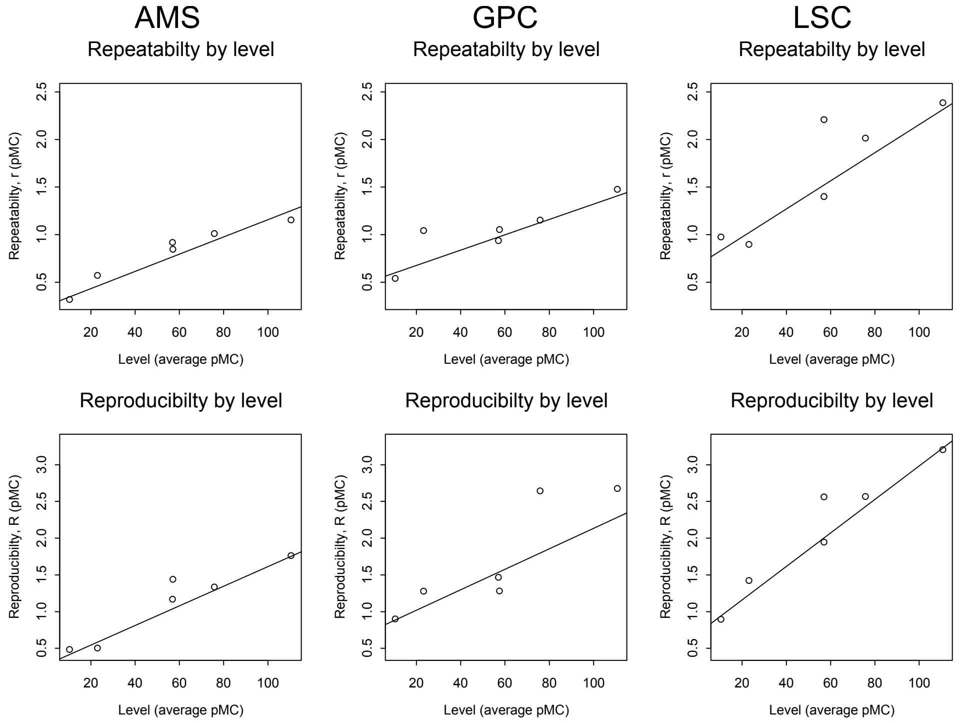

Figure 5.13 Repeatability and reproducibility for laboratory types

https://doi.org/10.1017/S0033822200032616 Published online by Cambridge University Press

5.6.5.2 Comments

Differences between the measurement techniques are observed. The AMS technique has lower reproducibility values compared to radiometric methods. LSC has higher repeatability values than the other techniques. Again, based on all the materials, the LSC results would be expected to be more varied than those from AMS or GPC laboratories.

5.6.6 How Can the Reliability Figures Be Used for Each Laboratory?

In essence, each laboratory may use its reliability figure to "test" whether it is sufficiently close to the consensus value for a reference material or standard.

Comparison with a reference value for a single laboratory makes use of *R*. If a single determination is performed by one laboratory under repeatability conditions and yields a value y^*, which is to be compared to the reference value m_0, then the critical difference (95%) between y^* and m_0 is given by:

$$CR = R / \sqrt{2}$$

If the absolute difference exceeds this critical difference, then the determination should be considered suspect and there may be an assignable cause that should be investigated. Assuming the reproducibility values given in Table 5.8, then for each of the samples, we can calculate the critical difference (CR) for a number (n) of independent determinations.

Table 5.11 Critical differences for each sample

Number of determinations	C	E	DF	I	H	GJ
1	0.56	0.81	1.13	1.45	1.54	1.87
2	0.43	0.70	0.97	1.19	1.33	1.62
3	0.37	0.65	0.91	1.09	1.26	1.53
4	0.34	0.63	0.88	1.04	1.22	1.48
5	0.32	0.62	0.86	1.01	1.19	1.45
6	0.31	0.61	0.85	0.99	1.18	1.43
7	0.30	0.60	0.84	0.97	1.16	1.42
8	0.29	0.60	0.83	0.96	1.15	1.40

Similar calculations can also be performed for AMS, GPC, and LSC techniques separately.

Table 5.12a Critical differences for each sample for AMS laboratories

Number of determinations	C	E	DF	I	H	GJ
1	0.34	0.36	0.83	1.02	0.94	1.25
2	0.30	0.21	0.69	0.92	0.80	1.10
3	0.29	0.13	0.63	0.89	0.74	1.05
4	0.28	0.06	0.61	0.88	0.71	1.03

Table 5.12b Critical differences for each sample for GPC laboratories

Number of determinations	C	E	DF	I	H	GJ
1	0.64	0.90	1.04	0.90	1.87	1.89
2	0.58	0.74	0.92	0.74	1.78	1.74
3	0.55	0.67	0.88	0.67	1.75	1.69
4	0.54	0.64	0.86	0.63	1.73	1.66

https://doi.org/10.1017/S0033822200032616 Published online by Cambridge University Press

Table 5.12c Critical differences for each sample for LSC laboratories

Number of determinations	C	E	DF	I	H	GJ
1	0.63	1.01	1.38	1.81	1.82	2.27
2	0.40	0.90	1.18	1.43	1.51	1.93
3	0.29	0.86	1.11	1.29	1.39	1.80
4	0.21	0.84	1.08	1.20	1.33	1.73

5.6.6.1 Comments and Conclusions

The critical differences decrease as the number of determinations increases; thus, the overall precision of the measurement increases as would be expected. The critical differences are a function of the material activity (an almost linear relation). We can also observe differences among the 3 measurement techniques, with AMS being more precise (given the realistic possibility of multiple determinations) than either GPC or LSC.

5.7 CONCLUSIONS

This section has mainly focused on the duplicate samples and their relationship to precision (taking account of the laboratory quoted error). On average, the difference in duplicate samples is zero, but there is some suggestion that the variation in the differences is greater than would be expected given the laboratory quoted errors. There is also a strong indication that the duplicate variation is considerably greater than would be expected in the near background Samples A and B.

Estimation of reproducibility and repeatability coefficients for firstly, the duplicate samples, and then for all materials, shows that the repeatability (measurements made under identical conditions in one laboratory) is a function of the sample activity and that the repeatability is better for the AMS technique than for the radiometric techniques. Reproducibility shows a similar pattern. Calculation of critical differences indicate that for a single determination, a relative difference from the consensus for Sample C greater than 0.05 pMC; for Sample E of 0.033 pMC; D, F, and I of 0.02 pMC; H of 0.02 pMC; and 0.017 pMC for GJ, would indicate that the measurement is aberrant.

This analysis does, however, make the assumption that the "average" quoted error is the same for all laboratories, which is clearly not the case.

https://doi.org/10.1017/S0033822200032616 Published online by Cambridge University Press