Supporting Information for Inferring Alignments I: Exploring the Accuracy and Precision of Two Statistical Approaches

Fabio Silva

Faculty of Humanities & Performing Arts, University of Wales Trinity Saint David; and IPHES, Institut Català de Paleoecologia Humana i Evolució Social, Àrea de Prehistòria, Universitat Rovira i Virgili (URV)
f.silva@uwtsd.ac.uk

Introduction

This document provides supporting information on the analysis of the main text. As its details are quite technical or tangential to the argument of the main text, it was decided they would be best presented in an online-only supporting text.

Code

The R code used for these simulations, including both the Alignment Model and the implementation of the curvigram and maximum likelihood (ML) methods as well as the resulting fitted equations (see below), is available in the author's GitHub page (https://github.com/f-silva-archaeo/InferAlignments1).

Results

Precision of the Curvigram Method

To estimate the precision of the curvigram method, one has to derive an equation from the simulated data (unlike for the ML approach, where an algebraic equation can be derived from first principles, see below). To do this, and because of the extra parameter in this approach, we have simulated 10 x 20 x 15 parameter combinations with 10,000 Monte Carlo iterations for each combination. The measurement uncertainty was varied from 1° to 10°, the deviation from 1° to 20° and the number of sites could take one of

JSA 3.1 (2017) S1–S10
https://doi.org/10.1558/jsa.34234

ISSN (print) 2055-348X
ISSN (online) 2055-3498

the following values: 5, 10, 20, 50, 75, 100, 150, 200, 250, 300, 350, 400, 450, 500 or 1000. Figure SF1 shows how the results vary for two values of the measurement uncertainty different than that shown in figure 8 of the main text.

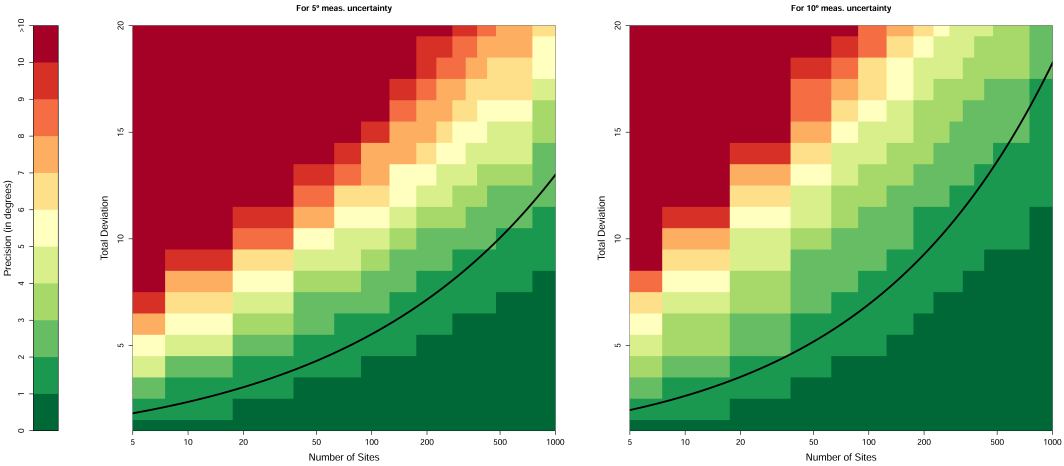

FIGURE SF1. Inferential precision for the curvigram single target scenario for two levels of measurement uncertainty and varying levels of deviation from target and number of surveyed sites. These figures are directly comparable to Figure 8 in the main text. The black lines are a fit to the 2° precision values.

To be able to predict the precision of the method when using it on an empirical dataset, an equation was fitted to this data using R (2016). Since one is interested in an accurate emulation of the results (as opposed to a parsimonious model), and there is no theoretical background to suggest a particular relationship in this case, the fit was done to a maximum polynomial degree of two for each variable with possibility of interactions between them. An analysis of variance (ANOVA) test was then conducted in order to choose the best explanatory model, which turned out to be the most complex one. The resulting equation is quite complex, with 27 coefficients, but the expected values correlate well with the results of the simulation (r-squared of 0.9969), meaning that the equation provides an accurate estimate for the precision of the curvigram method. The equation has the following form:

$$\begin{aligned} prec = {} & \emptyset + A\sigma + B\sigma^2 + Clog(N) + D\log(N)^2 + E\delta + F\delta^2 + G\sigma\log(N) + H\sigma^2\log(N) + \\ & + I\sigma\log(N)^2 + J\sigma^2\log(N)^2 + K\sigma\delta + L\sigma^2\delta + M\sigma\delta^2 + N\sigma^2\delta^2 + Olog(N)\delta + \\ & + P\log(N)^2\delta + Qlog(N)\delta^2 + R\log(N)^2\delta^2 + S\sigma\log(N)\delta + T\sigma^2\log(N)\delta + \\ & + U\sigma\log(N)^2\delta + V\sigma^2\log(N)^2\delta + W\sigma\log(N)\delta^2 + X\sigma^2\log(N)\delta^2 + \\ & + Y\sigma\log(N)^2\delta^2 + Z\sigma^2\log(N)^2\delta^2\,, \end{aligned} \qquad \text{(SE1)}$$

where σ is the the standard deviation in the dataset which, in the main text, we have called *total deviation*, δ is the measurement uncertainty and $\log(N)$ is the natural logarithm of N, the number of archaeological sites in the dataset (i.e. the sample size). The estimates for the lettered coefficients are given in table ST1.

TABLE ST1. Coefficients for equation SE1, estimated via a multivariate regression.

Coefficient	Estimate	Std Error
Ø	2.568E-02	5.623E-01
A	2.391E+00	1.233E-01
B	-3.620E-02	5.704E-03
C	-5.822E-01	2.806E-01
D	5.815E-02	3.252E-02
E	-1.004E+00	2.348E-01
F	7.711E-02	2.081E-02
G	-2.673E-01	6.154E-02
H	5.132E-03	2.846E+00
I	3.285E-03	7.131E-03
J	-2.257E-05	3.299E-04
K	2.561E-01	5.150E-02
L	-1.179E-02	2.382E-03
M	-2.821E-02	4.563E-03
N	1.535E-03	2.111E-04
O	3.493E-01	1.172E-01
P	-1.991E-02	1.358E-02
Q	-1.643E-02	1.038E-02
R	1.099E-05	1.203E-03
S	-1.915E-01	2.570E-02
T	9.331E-03	1.189E-03
U	1.748E-02	2.978E-03
V	-9.655E-04	1.378E-04
W	1.334E-02	2.277E-03
X	-8.288E-04	1.053E-04
Y	-9.556E-04	2.639E-04
Z	7.335E-05	1.221E-05

As this equation is quite complex, a simpler relationship that could provide a measure of the minimum number of sites required to have good precision was sought. Fitting the results of the simulation for a precision of 2° yielded the following power-law relation:

$$\sigma > N^x, \qquad \text{(SE2)}$$

where *x*, the power-law coefficient, varies for different values of the measurement uncertainty. This relationship can be obtained from the simulated data by fitting the value of *x* versus the values of measurement uncertainty used. The relationship turns out to be log-linear (r-squared of 0.996902) as shown in figure SF2.

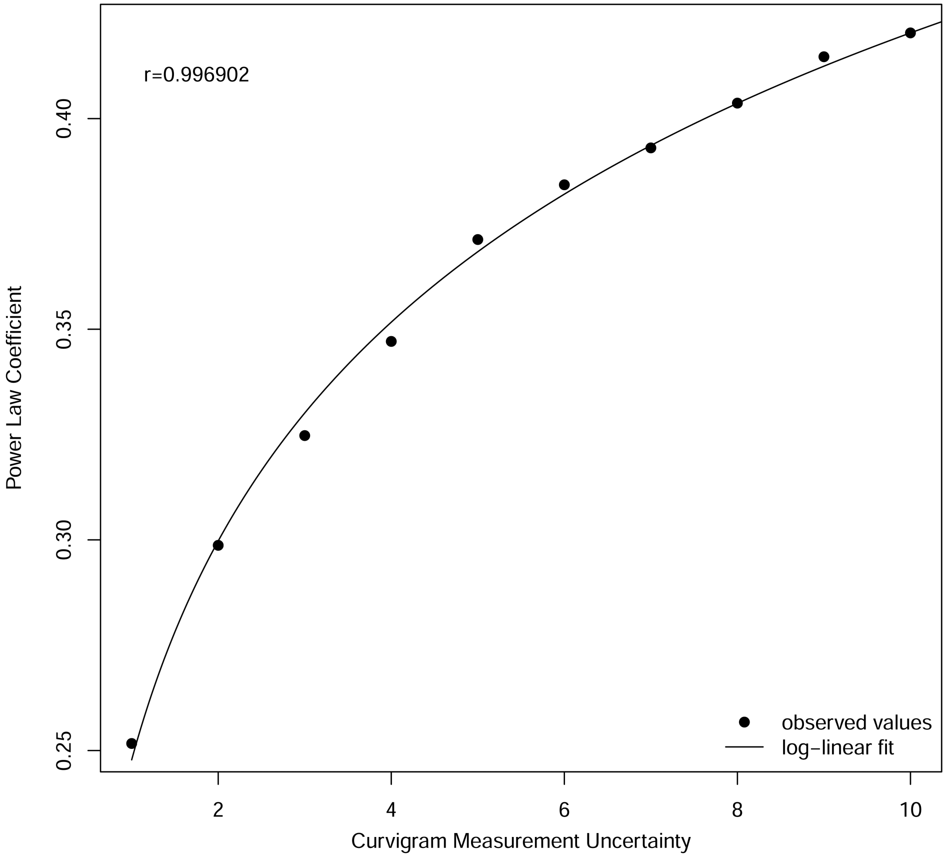

FIGURE SF2. Evolution of the coefficient of the power-law relation SE2 for varying values of measurement uncertainty.

This yields the following equation for estimating the minimum number of sites required to ensure that the curvigram method has a precision of 2° or better:

$$N > \sigma^{\frac{1}{a+b.log(\delta)}}, \qquad \text{(SE3)}$$

where the coefficients are given in table ST2.

TABLE ST2. Coefficients for equation SE3, estimated via linear regression.

Coefficient	Estimate	Std Error
a	0.247799	0.002456
b	0.074932	0.001477

Closer scrutiny of this relation further highlights the point made in the main text that the measurement uncertainty should not be lower than the deviation present in the data. Figure SF3 plots equation SE3 for the minimum number of sites necessary to have precision of 2° or better for five different scenarios of varying deviation (black curves). The figure makes it clear that, for ratios above unity (to the right of the vertical blue line), that is for curvigram analyses where the measurement uncertainty is larger than the deviation present in the data, the number of sites required to achieve high precision is quite low. Conversely, for ratios below unity (i.e. to the left of the vertical blue line), the minimum number of sites increases dramatically, emphasising the need to use an uncertainty that is larger than the deviation in the dataset.

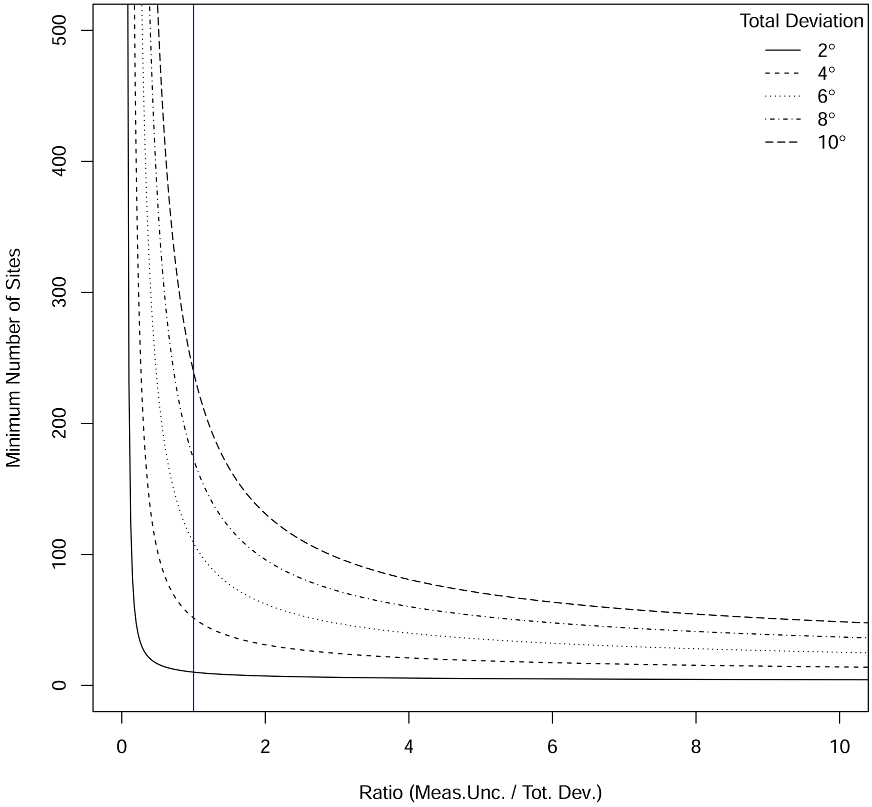

FIGURE SF3. Minimum number of sites required to ensure a high precision versus the ratio between the curvigram measurement uncertainty and the total deviation in the data for varying values of total deviation (black curves). The vertical blue line marks a ratio equal to unity, i.e. when measurement uncertainty and total deviation are the same.

Measurement Uncertainty vs Bandwidth

Kernel Density Estimators (KDE) use the bandwidth parameter in order to smooth the resulting distribution in such a way as to more accurately recover the limiting distribution. This is a free parameter for which there is no theoretical value and which is often estimated. Most standard estimates (Jones *et al.* 1996) agree that the bandwidth should be inversely proportional to the sample size, meaning that it should be larger for small number of samples (and therefore providing more smoothing) and smaller for large sample sizes. We are now in a position to compare the ideal value for the measurement uncertainty and that given by KDE bandwidth estimators.

Most estimators, such as MISE and AMISE (Jones *et al.* 1996), are quite complicated to implement and data-specific. We therefore look at one of the simplest, Silverman's rule-of thumb estimator (Silverman 1986), which is appropriate for situations where the limiting distribution is a Gaussian curve, as is the case with the Alignment Model implement in this paper. Silverman's rule of thumb for the KDE bandwidth parameter (*bw*) is of the form:

$$bw = 1.06.\sigma.N^{-1/5}, \qquad \text{(SE4)}$$

whereas equation SE3 can be rewritten to express the minimum measurement uncertainty that one should use in the curvigram method to ensure that the estimate will be precise to within 2°. This effectively constitutes an estimator for KDE bandwidth which, for the single target scenario explored in this paper, ensures that the inferred precision is of 2° or better. This new estimator is:

$$\delta > e^{\frac{1}{b}\cdot\left[\log\sigma / \log N - a\right]}. \qquad \text{(SE5)}$$

Figure SF4 compares these two estimators for varying numbers of sites and values of total deviation. We can see that, for any given total deviation value, Silverman's rule-of-thumb underestimates the bandwidth up to a certain number of sites, after which the bandwidth is above the minimum measurement uncertainty to ensure high precision, making it optimal. This doesn't prevent other, more powerful, bandwidth estimators from performing better. This will be easily checked by future scholars by comparing the estimators with equation SE5 above.

Precision of the Maximum Likelihood Method (ML)

For the single target scenario, where all measurements have the same uncertainty, the ML estimate is equal to the mean (Taylor 1997, 97) which is given by:

$$x_{mean} = \frac{\sum x_i}{N}, \qquad \text{(SE6)}$$

where x_i is an individual measurement and N is the number of measurements (the sample

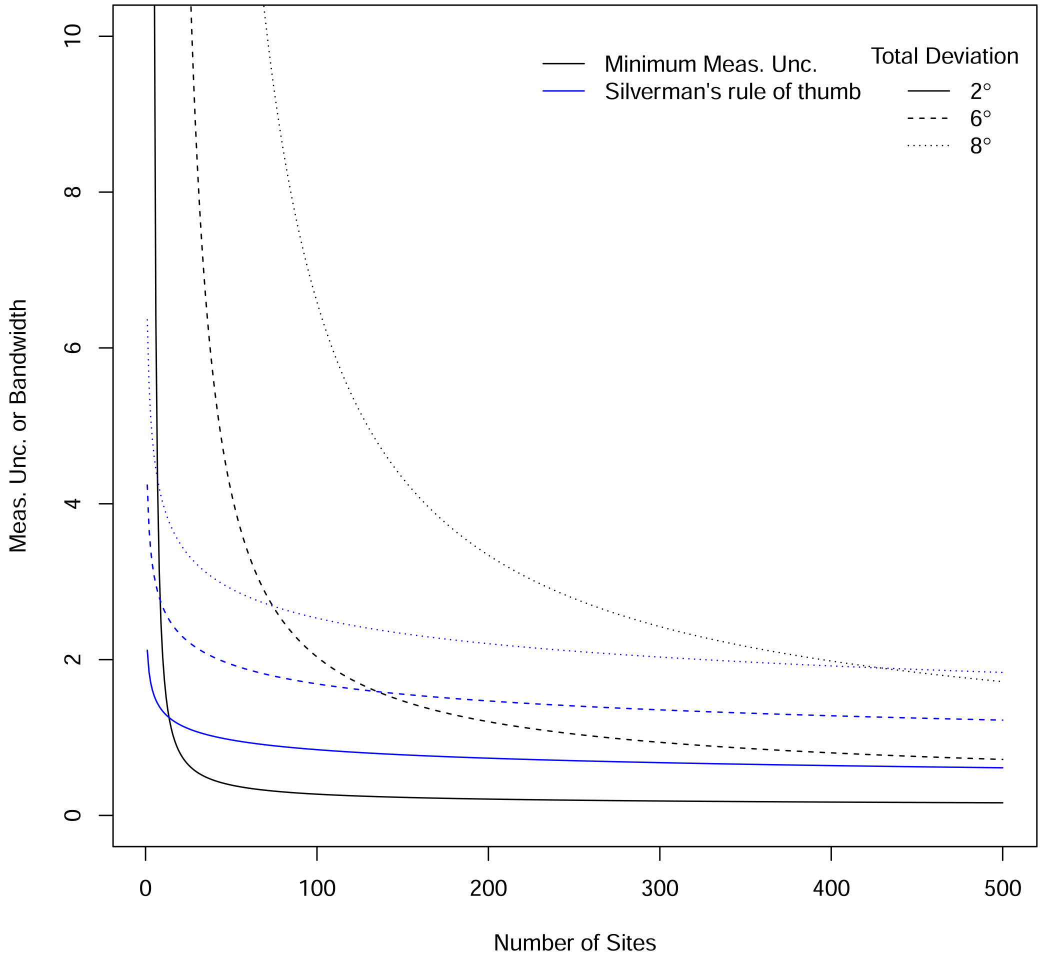

FIGURE SF4. Minimum Measurement Uncertainty to ensure precision of 2° or better (black curves) versus Silverman's rule-of-thumb estimator for the KDE bandwidth parameter (blue curves) for varying number of sites and values of total deviation.

size). This mean has an associated error that is given by the so-called *standard deviation of the mean* (Taylor 1997, 102):

$$\sigma_{mean} = \frac{\sigma}{\sqrt{N}} \quad , \qquad \text{(SE7)}$$

where σ is the standard deviation in the dataset which, in the main text, we have called *total deviation*. The theoretical precision for the ML method, therefore, is associated to this standard deviation of the mean. To work at 95% confidence level, the precision is given by 1.96 standard deviations:

$$precision = 1.96\ \sigma_{mean}\ , \text{ at 95\% confidence} \qquad \text{(SE8)}$$

It then becomes trivial to estimate how many samples (i.e. how many archaeological sites) are needed to achieve a given precision level, *P*. The equation is derived as follows:

$$1.96\,\sigma_{mean} \leq P \qquad ,$$

$$1.96\frac{\sigma}{\sqrt{N}} \leq P \qquad ,$$

$$N \geq \left(\frac{1.96}{P}\sigma\right)^2 \qquad . \qquad \text{(SE9)}$$

Therefore, for a minimum precision of two degrees (P=2°) we get:

$$N \geq 0.98\,\sigma^2 \approx \sigma^2 \quad , \qquad \text{(SE10)}$$

which is the same as equation E2 in the main text. Alternatively, a combination of equations SE7 and SE8 allows one to estimate the precision for any empirical dataset as:

$$precision = 1.96\,\sigma_{mean} = 1.96\frac{\breve{}}{\sqrt{N}} \quad . \qquad \text{(SE11)}$$

To see how the theoretical expectation for the precision (equation SE11) compares with the results of the analysis of the Monte Carlo simulations, the expected value has been plotted against the obtained values in figure SF5 below. They closely follow the identity line (in blue) and produce a very high correlation coefficient of 0.99795, therefore demonstrating that the algebraically-derived equation is a good predictor for the precision of the ML method.

Discussion

Equations SE3 and SE10 give the minimum number of sites required for each of the methods to have a precision of 2° and, therefore, they are directly comparable – the key difference being the power-law coefficient. Figure SF6 compares the values of this coefficient for different measurement uncertainties. It makes clear that, for the (low) levels of curvigram uncertainty typically considered by archaeoastronomers, the coefficient is always larger than that of the ML method, meaning that the curvigram method will require a much larger sample size to ensure high precision. The two methods have the same coefficient at a value of measurement uncertainty given by:

$$mes.unc. = e^{\frac{0.5-a}{b}} = 28.95467 \quad , \qquad \text{(SE12)}$$

which is too high for any practical purposes.

This short analysis complements the discussion in the main text and shows that the curvigram method is highly sensitive to the measurement uncertainty. If the latter is underestimated, then the method becomes very imprecise, requiring considerably more structures to be surveyed and analysed in order to reach the same levels of precision of the ML method.

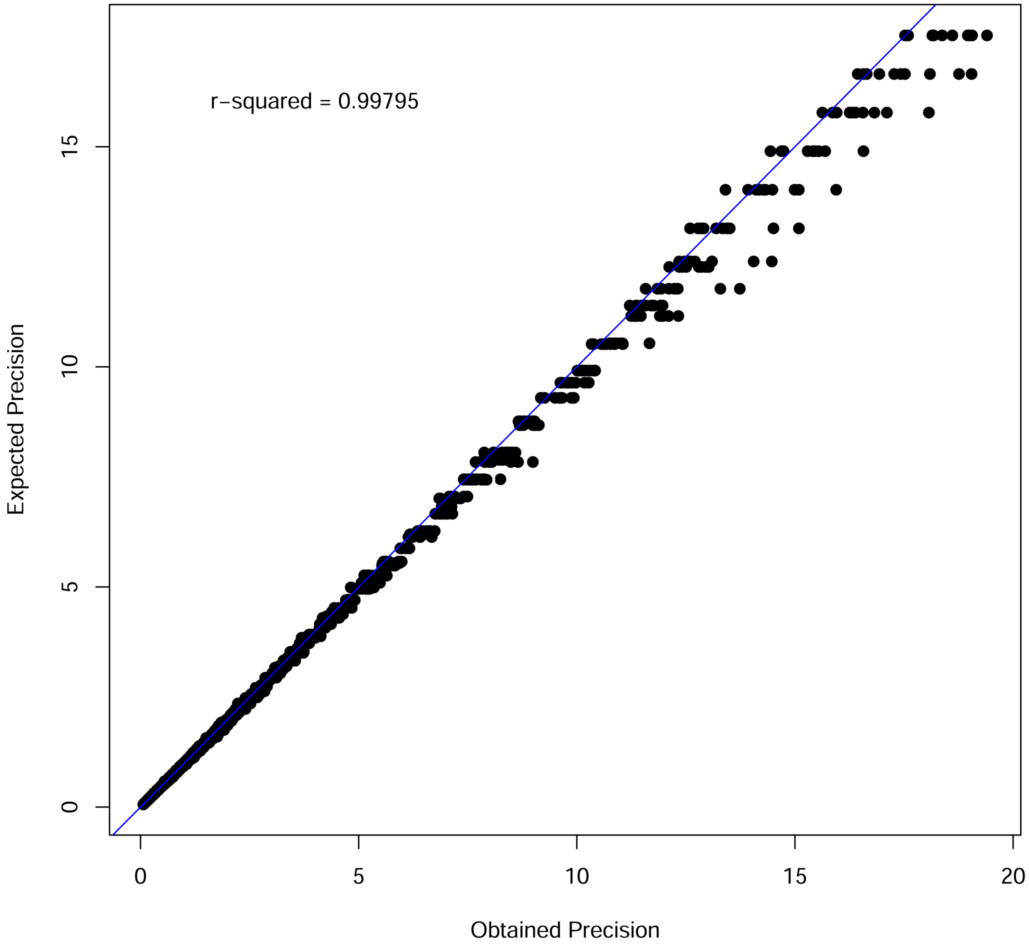

FIGURE SF5. Comparison of the expected (theoretical) and obtained (simulated) precision of the ML method, also showing their correlation coefficient and the identity line (in solid blue).

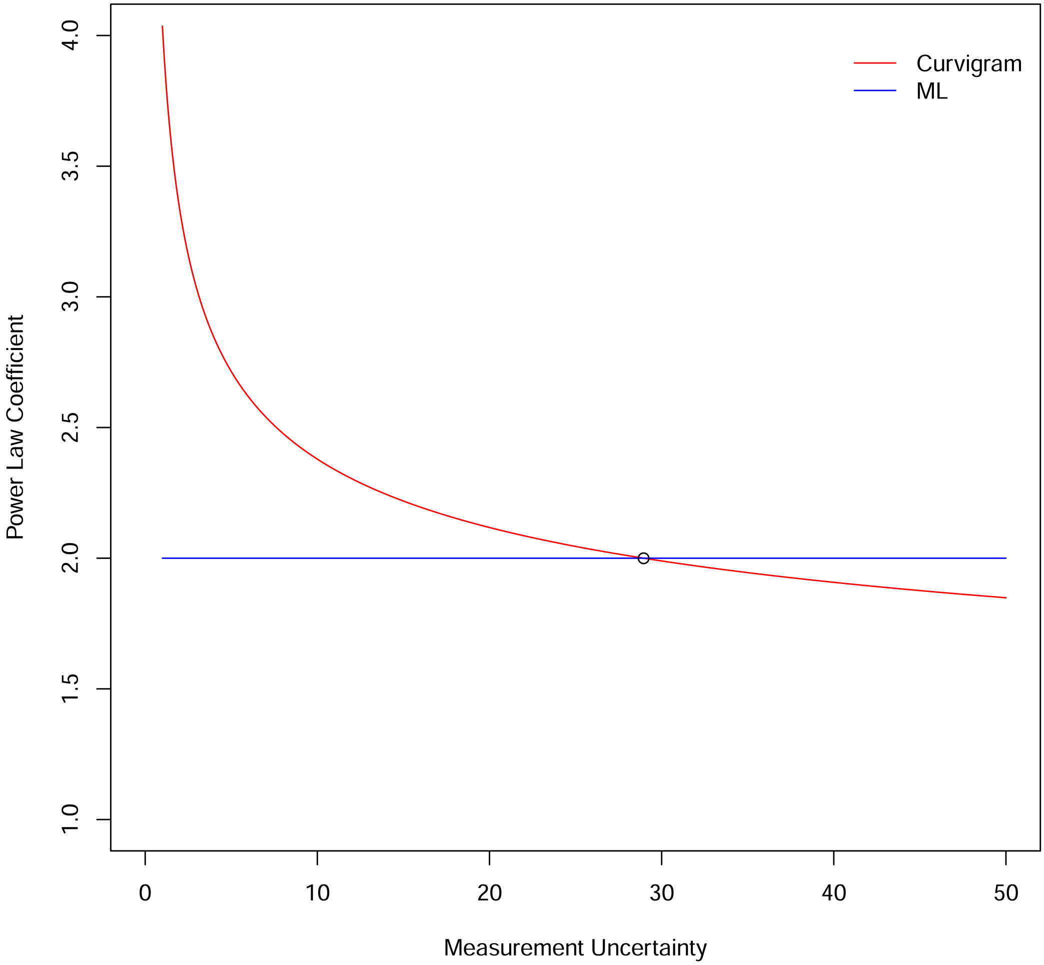

FIGURE SF6. Power-law coefficient of equations SE3 (red curve) and SE10 (blue line) for the curvigram and ML methods, respectively. The small circle marks the value of measurement uncertainty for which both methods have the same coefficient and, therefore, require the same number of sites to reach the same level of precision.

References

Jones, M.C., J. S. Marron and S. J. Sheather, 1996. "A Brief Survey of Bandwidth Selection for Density Estimation". *Journal of the American Statistical Association* 91 (433): 401–407. https://doi.org/10.1080/01621459.1996.10476701

Silverman, B. W., 1986. *Density Estimation for Statistics and Data Analysis*. London: Chapman and Hall / CRC. https://doi.org/10.1007/978-1-4899-3324-9

Taylor, J. R., 1997. *An Introduction to Error Analysis: The Study of Uncertainties in Physical Measurements*. Sausalito, CA: University Science Books.