Published in *Journal of Skyscape Archaeology* 3(1): 93-111. DOI: 10.1558/jsa.31958

Inferring Alignments I: exploring the accuracy and precision of two statistical approaches

Fabio Silva

Abstract
Using computer simulations, this paper explores and quantifies the accuracy and precision of two approaches to the statistical inference of the most-likely targets of a set of structural orientations. It discusses the curvigram method of wide currency in archaeoastronomy (also known as Kernel Density Estimation or Summed Probability Densities), and introduces the largely unused maximum likelihood method, that is quite popular in other fields. The analysis of their accuracy and precision is done for a scenario with a single target, from which resulted equations that can be used to estimate the minimum number of surveyed structures to ensure a high-precision statistical inference. Two fundamental observations are also made: that, although both approaches are quite accurate, the maximum likelihood approach is considerably more precise than the curvigram approach; and that underestimating measurement uncertainty severely undermines the precision of the curvigram method. Finally, the implications of these observations for past, present and future archaeoastronomical research, is discussed.

Keywords: Statistics, Inference, Likelihood, Curvigram, Accuracy, Precision

Introduction: inference and simulation in skyscape research

Pattern recognition is a cornerstone of the study of the orientation of similar archaeological structures, whether one is primarily looking at the landscape or the skyscape. The logic behind it is quite sound: a random distribution of orientations would feature no patterns whatsoever, therefore a pattern in structural orientation betrays intent on the part of their builders. This approach has had wide application both from a qualitative analysis point of view (e.g. Tilley 1997; Cummings et al 2002; Sims 2009) as well as quantitative (e.g. Ruggles 1999, Pimenta et al 2015, Prendergast 2011, Bevan and Lake 2013).

The goal is, typically, to identify a significant deviation from random chance and, as the orientation data is quantitative, statistical inference is particularly useful here. In the field of Statistics, inference is defined as "the process of coming to some conclusion about a population based on a sample" (Clapham and Nicholson 2014: 239). The population is the collection of data "that would be obtained if the number of measurements become[sic] infinitely large" (Taylor 1997, 121). It is often impossible and/or undesirable to take such large number of measurements and therefore, statistical inference, through the analyses of the available sample, aims to estimate what is known as the *limiting distribution*, a theoretical, unrealistic and even ideal, construct that contains all the relevant information about the population. This stands in sharp contrast with descriptive statistics which is "concerned with describing the basic statistical features of a set of observations" (Clapham and Nicholson 2014: 127), rather than infer something about the population from which the observations were taken.

Statistical inference is appropriate to analyse structural orientation data since the set of measurements is, by its very nature, a sample of a larger set: in archaeology, and particularly in prehistory, one hardly ever has the entire set of, say, all the stone circles built at a particular time period by a particular culture – many would have been lost in the intervening millennia for a variety of reasons, whereas on other cases a sufficiently large number of similar structures was never built in the first instance. This raises the one limitation of the application of this sort of statistics: it requires, or hopes for, a large number of measurements: the application of inferential techniques on orientation data of a small number of structures is inherently flawed. A pattern of one is no pattern at all, and this has even led some archaeoastronomers to discount the analysis of singular structures (e.g. Hoskin 2002; Belmonte 2006). A pattern of two in a sample of five, might not be sufficient to discount the hypothesis that it might just be a coincidence. On the other hand, a pattern covering one hundred out of 110 structures seems significant enough to warrant closer inspection.

This has, in the past, led to two schools of thought within archaeoastronomy. The followers of the 'green' school, typically interested in the orientations of European megaliths, would survey hundreds, if not thousands, of similar structures and apply statistical techniques to describe their data and infer possible celestial targets (Heggie 1982). The followers of the 'brown' school, however, typically those interested in Amerindian cultures and Mesoamerican structures, would instead recur to the ethnographic and historical record in order to substantiate any claims for structural alignments, typically of single structures or complexes (Aveni 1982). This gap is, however, narrowing as both schools are coming closer together, with American archaeoastronomers now using statistical analyses of larger datasets (e.g. Sprajc 2015; González-García and Sprajc 2016), and European archaeoastronomers starting to take the wider archaeological and even ethnographical records into account (e.g. Henty 2014, Silva 2015).

Nevertheless, as recently highlighted by Ruggles (2015, 420), the exploration and formalization of inferential techniques in archaeoastronomy was mostly confined to the field's early development in the seventies and eighties. These premature explorations focused almost entirely on the issue of statistical significance (e.g. Freeman and Elmore 1979) whereas other key fundamental questions were left unanswered or, worse, unasked.

Inferential uncertainty, accuracy and precision

Significance testing is very important since one needs to be confident that there really is a pattern in the data that cannot be explained by mere random processes. However, significance itself does not vouch for the accuracy of the inferred target. This paper therefore aims to add to the previous literature by focusing on this neglected yet fundamental issue: what is the uncertainty in the values inferred by the statistical methods we routinely employ to identify the targets of structural alignments? In other words, how accurate and precise are these methods? Despite this issue being largely recognized as of importance, it has never been studied in a systematic manner and is universally neglected in the archaeoastronomical literature – see for example its lack of mention in two recent textbook volumes (Ruggles 2015, Magli 2016).

This is a particularly important question since today's archaeoastronomers tend to "rely more in qualitative assessment of the declination distributions" (Ruggles 2015, 420), identifying

significant frequency peaks in histograms and curvigrams and matching them to the nearest celestial object or event. However, these are the tools of descriptive, rather than inferential statistics: they describe the data available but when it comes to inferring something from them, other tools can be more accurate and precise. But since the issues of accuracy and precision have been ignored the application of these methods might have been (mis)leading researchers into erroneous interpretations. Some of these interpretations might have been unquestioningly propagated in the literature, leading countless other scholars astray, particularly when independent evidence is unavailable.

The importance of quantifying and reporting inferential uncertainty is illustrated in a thought experiment where two experts are invited to try to ascertain whether a crown is made of 18-karat gold or of a cheap alloy (Taylor 1997: 5-6). Since this cannot be measured directly, it has to be inferred from another measurement. In this case, the density of the crown is a good proxy since 18-karat gold has a known density of 15.5 gram/cm^3 whereas that of the cheap alloy being considered is 13.8 gram/cm^3. Figure 1 shows the density values measured by the two experts. George first reports his best estimate to be 15 gram/cm^3, with no mention of the uncertainty surrounding this value. Based on its proximity to the density of 18-karat gold, Bob would infer that the crown must be made of gold. However, when pressed about the level of uncertainty in his measurement he reports a range of values that encompasses both the value of gold and the cheap alloy, implying that, based on this measurement alone, one cannot infer with any certainty whether the crown is made of gold or the cheaper alloy. Martha, the second expert, reports her measurement to be 13.9±0.2 gram/cm^3, meaning that she is pretty confident that the crown's density lies between 13.7 and 14.1 gram/cm^3. This figure is not only consistent with George's measurement (i.e. there is an overlap) but the uncertainty in Martha's measurement is significantly lower (figure 1). In fact, the range of likely values now excludes the possibility that the crown is made of gold, meaning one is in a better position to infer the crown to be made of the cheaper alloy. Note how much of a difference having the uncertainty made: by ignoring or underestimating the uncertainty in George' measurement, one would have erroneously inferred the crown to be made of gold!

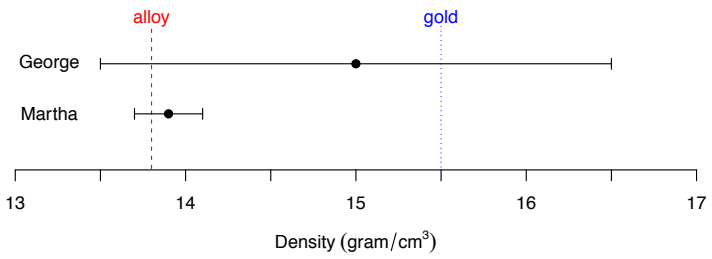

Figure 1 – Two measurements of the density of a metal crown. The black dots show George and Martha's best estimates, whereas the horizontal bars show their uncertainties. George's uncertainty is so large that both gold (blue dotted line) and a cheaper alloy (red dashed line) fall within its range, making it impossible to infer which metal the crown is made from. On the other hand, Martha's measurement clearly shows that the crown is made from a cheaper alloy.

Taylor's thought experiment can be readily extended to the case of structural orientations: by ignoring uncertainty and, instead, inferring targets based on the qualitative assessment that the obtained value is closer to one potential target than to another (as in George's first assessment), one might be inferring the wrong conclusions from the data. Furthermore, an alternative technique, like Martha's, can result in lower uncertainties and therefore in a better inference.

The issue of inferential uncertainty requires the introduction of the two complementary concepts of *accuracy* and *precision*. Accuracy relates to how close the inferred values are to the true value, whereas precision relates to how repeated measurements under the same conditions would show the same results. The two concepts are easily confused but, in truth, they are unrelated: a particular technique might have low accuracy but high precision, or vice-versa. This is made plain when one considers the archery analogy, where precision relates to whether arrows form a tight cluster or are scattered in the target, and accuracy relates to how close to the bullseye the centre of the cluster is. Yet, when one is applying an inferential method to an empirical dataset, where one does not independently know what is the true value one is trying to infer, nor do we have multiple datasets of the same archaeological structures available, it is impossible to know the method's accuracy or precision a priori. All one can know is *how likely* the method is to be accurate and/or precise, and quantify their *expected* values. This can be done using an approach not dissimilar to that used by natural and health scientists. In these research arenas, scholars can quantify the accuracy (or effectiveness) of a particular treatment by running a large number of experiments where extraneous factors are controlled for. By applying the treatment to a large sets of patients one can assess its effectiveness by counting how many patients felt better after the treatment.

In archaeology, however, one doesn't have the benefit of having another thousand sets of similar monuments lying around which could be used for controlled trials. Instead, computer modelling and simulation are being used to this effect (e.g. Barcelo et al 2015; Lake 2014; Silva et al 2015). For the particular type of problem that this paper addresses, a computer model can be created to mimic the orientation of a set of structures, where parameters such as the intended target and level of uncertainty can be controlled, but the noise added to mimic that uncertainty is stochastic (i.e. random). This can be done hundreds, thousands or even tens of thousands of times, always producing a different dataset based on the same parameters, mimicking the natural scientist's multiple controlled experiments. This approach is known as the *Monte Carlo method*, named after the famous casino in Monaco (Fishman 1995). The statistician can then apply his methods to the simulated datasets created by the Monte Carlo algorithm, compare the method's inferred results with the known parameters and therefore quantify the expected accuracy and precision of the statistical method.

This paper takes exactly this approach. Firstly, it establishes a quantitative model for structural alignments that is then used, in conjunction with the Monte Carlo method, to assess the accuracy and precision of two approaches to statistical inference: one widely used by archaeoastronomers, the other more common in other fields. This is done for a scenario involving a single intentional target of fixed declination, but varying noise levels. Accuracy and precision are expected to vary for different datasets since these will present different conditions, such as different number of archaeological sites. It is therefore important to explore how they vary for different combinations of the model's parameters. By doing this

one will also be able to answer an important corollary question such as what is the minimum number of archaeological sites required to ensure that targets are inferred with high precision.

Methods

The Alignment Model

In order to simulate a set of structural alignments one needs to construct a reasonable model that mimics the process that moved from the limiting distribution – i.e. the intentional celestial target(s) – to the empirical data – i.e. the field measurements. Without any loss of generality this can be conceived of as a three stage process (depicted in figure 2).

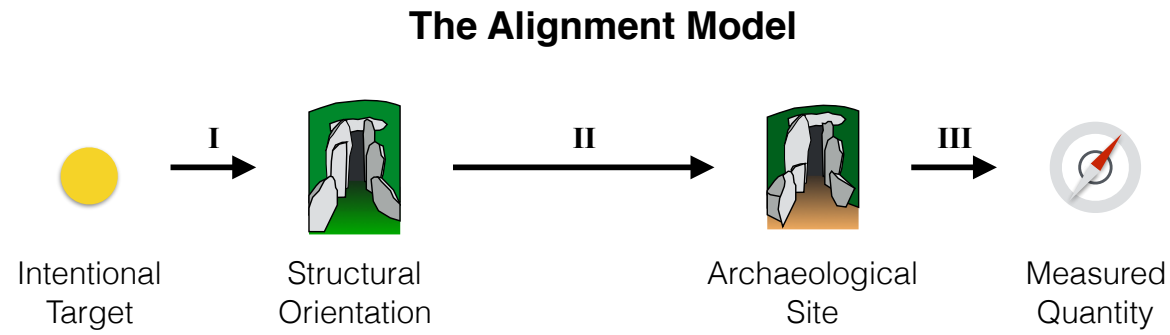

Figure 2 – The model for structural alignments. From a past society identifying the orientation of the intentional target and encoding it into the architecture of a structure (stage I), through the effects of time and erosion on said structure (stage II), to the measurement of its orientation (stage III). These three stages involve the addition of noise or uncertainty to the intended alignment.

Stage I is the process by which the past society observes and identifies a topographic or celestial object as a target for the alignment, followed by its encoding into the architecture of a structure. This process was not necessarily accurate, as it would have involved some degree of error or uncertainty. This could be related to the use of imprecise instruments to mark the intended orientation, a more general lack of interest in precision, the fact that an error was made, or the choice of a celestial event that does not always occur on the same spot of the horizon, such as Equinoctial Full Moons (Silva and Pimenta 2012), the Summer Full Moon (e.g. Ruggles 1999), sunrise over a prolonged period (e.g. Hoskin 2002) or the appearance of a comet.

Stage II corresponds to the processes by whose means time, erosion and other historic effects, take their toll on the structure in question, transforming it into the archaeological site (often ruinous) that we have today. This too will be a source of uncertainty: as part, or all, of the structure might be ruined. It might later be surveyed and even reconstructed by archaeologists, but this will further add some degree of uncertainty.

Stage III is the surveying of the site for its orientation and here too, the choice of surveying methodology and instrument adds uncertainty to the measured quantity.

Such a model has five parameters: the orientation of the intentional target, the number of structures that were built in alignment with this target, the deviation in the structural orientation with respect to the target (i.e. the uncertainty introduced in stage I), the deviation in orientation due to weathering (stage II) and the deviation in the measurement of this orientation (stage III). Although measurement uncertainty can be estimated, and stage II uncertainty minimized by taking a careful look at excavation plans or actually doing the measurements when the site is being excavated, stage I uncertainty is harder to estimate without making some assumptions. One would have to answer the question of how precise would these alignments have to be for past societies. This has been hotly debated since the claims of high-precision prehistoric alignments by Thom (e.g. 1955, 1967) and is far from being resolved.

This model is similar to that used by Pimenta et al (2009), except that they combined Stage I and Stage II uncertainties. However, since the different uncertainties are, in general, unknown we can combine all of them into a single *total deviation,* so as to fully simplify the model. This is only possible because the uncertainties related to the three stages are independent and one can safely assume them to be normally distributed. This simplified model contains only three parameters: the target orientation, the number of structures and the total deviation in the structural orientations with respect to the target. Unless there were any systematic errors present in any of the three stages outlined above, this simplified model is equivalent to the generalized model of figure 2.

The Alignment Model is then based on a normal limiting distribution, whose mean is equal to the target's true orientation, and whose standard deviation is equal to the total deviation (figure 3). This is to say that 68.2% of structures are expected to fall within one total deviation of the true target, and 95.4% within two total deviations. This is quite appropriate for fixed targets such as singular topographic features, or celestial objects/events that recur on the same orientation. The set of orientations output by the model is then taken, at random, from this distribution.

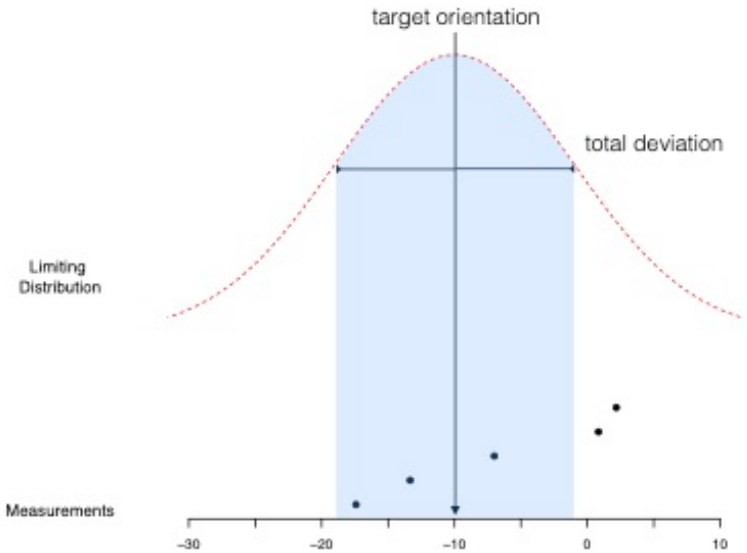

Figure 3 – A simulated set of five structural orientations (black dots at bottom) obtained from the simplified Alignment Model for a target orientation of -10° of declination (black vertical

arrow), and a total deviation of about 9° (blue shaded area), corresponding to the bell-shaped limiting distribution in red (top).

Implementing the model

To simulate this simplified model, a bespoke algorithm was constructed using R (2016). The algorithm allows one to control the model's parameters, namely: the orientation of the intended target (in this case a declination), the total deviation and the number of (simulated) archaeological sites supposed to be aligned to that target. The algorithm outputs an array containing the (simulated) measurements of orientation ready to be statistically analysed using the same methods that would be applied to an empirical dataset (see below).

Because the added uncertainty is random one cannot infer anything of value from a single simulated dataset. This is where the Monte Carlo method comes in: by repeating the process just outlined several thousand times, for the same parameter values, one can explore the range of possibilities and look for similarities and differences. When this is done one is left with thousands of possible datasets that represent the same original intention to align to a specific target, at a specific degree of deviation. One can then apply the chosen inference method to each dataset and see how well they recover the (known) value of the target orientation. For instance, if the true value is recovered in less than 5% of the simulated datasets, then it can be said that, given the used parameter values, the employed inference method is unlikely to be accurate.

The Curvigram Method

A popular inference method, that gained traction since Thom (1955, 1967) is to use what Ruggles named a curved histogram, or a curvigram (Ruggles 1981, 156; 2015, 418). They consist of summing the probability distributions of each measurement together, thus creating a distribution that represents the totality of the empirical data. This is an application of the Kernel Density Estimation (KDE) technique (e.g. Silverman 1986), typically using a Gaussian kernel, and where the bandwidth (this technique's parameter) and measurement uncertainty are inextricably linked. This sort of approach has also recently gained traction in archaeology where it is better know as the *summed probability distributions* method (SPD for short). Famously applied to the sum of calibrated probability distributions of radiocarbon dates, it has been argued that such an SPD of dates is a proxy for past population structure (e.g. Williams 2012; Shennan et al 2013; Stevens and Fuller 2012; Zahid et al 2015).

Any measurement consists of a best estimate (the value one reads on a measuring tape, for example) as well as the uncertainty associated with the measurement (for instance, the smallest scale division of the measuring tape). Instead of simply using the best estimate one can take the measurement uncertainty into account by constructing a probability distribution for the measurement (the dotted blue bell-shaped curves of figure 4). These represent the likelihood that the value matches the true orientation: the higher the curve the more likely the value of orientation is to match the measurement. These typically are of the Gaussian, or normal, variety so that the likelihood peaks at the best-estimate and quickly falls off outside of the uncertainty range. In this approach, the individual measurement distributions are summed, creating the curvigram (the solid blue curve in figure 4).

Published in *Journal of Skyscape Archaeology* 3(1): 93-111. DOI: 10.1558/jsa.31958

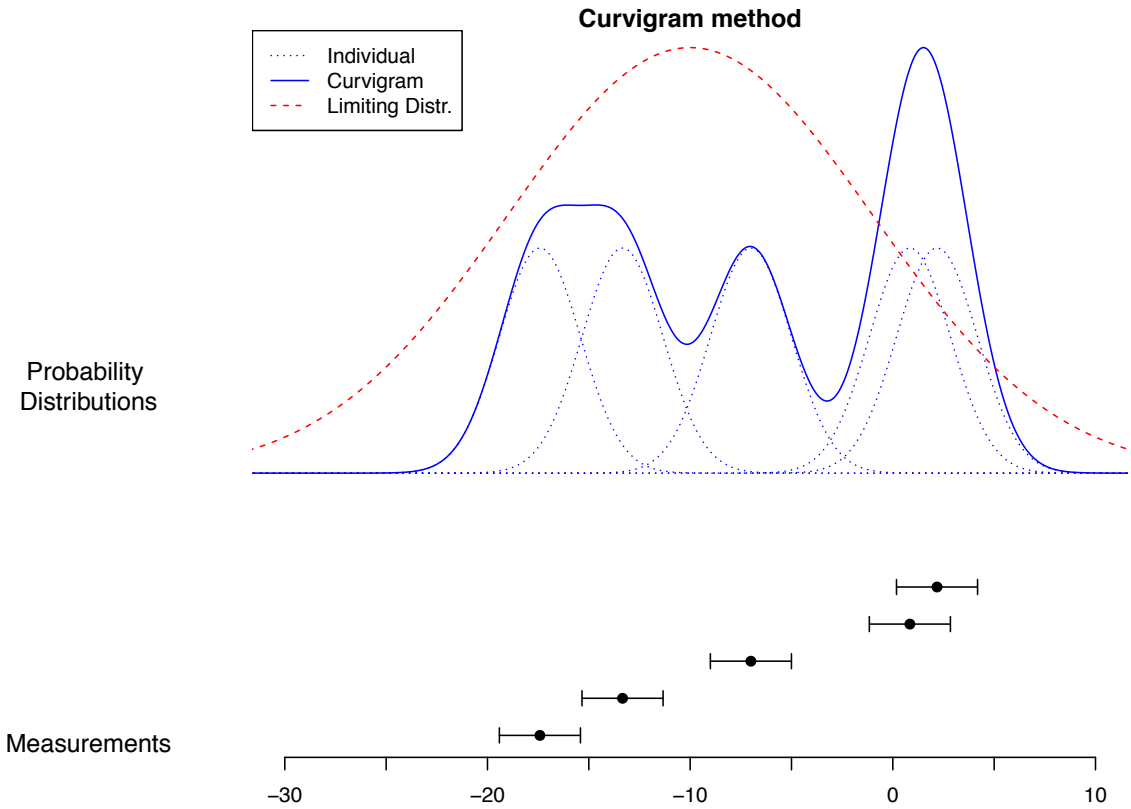

Figure 4 – The Curvigram method. The measurement's best estimate (dots at bottom) and associated uncertainty (error bars) are translated into individual bell-shaped probability distribution (blue dotted curves at top), which are then summed to create the curvigram or SPD (blue solid curve at top). Compare this to the limiting distribution from which the measurements are a sub-sample of (red dashed curve).

Like Kernel Density Estimation techniques, the Curvigram method is a bottoms-up approach: it tries to reconstruct the shape of a distribution from the empirical data. But the question it really is addressing is: "what does the distribution of the measurements looks like?" Just like histograms, the curvigram can only be expected to match the limiting distribution for large sample sizes (compare the obtained curvigram with the actual limiting distribution in figure 4). Nevertheless, archaeoastronomers routinely look at the peaks of the curvigram under the assumption that, since these represent the most frequent values, they should be close to the true orientation of the intended target(s). Unfortunately, this is not always the case: the curvigram in figure 4 peaks around a declination of -15°, whereas the true target is at -10°. Furthermore, curvigrams often display multiple peaks which can be interpreted not as insignificant artefacts of under-sampling, but as different targets, further leading the scholar astray. Finally, this approach does not have any theoretically-grounded estimate for inferential error, meaning that error margins are not estimated nor reported which, as discussed above, can be very damaging.

On the other hand, a clear advantage of this approach is that the uncertainty in each measurement is taken into account since the entire probability distribution is used, rather than simply the best estimate as is common when using mere histograms (e.g. Hoskin 2001). In fact, measurements with differing uncertainties can be brought into the same framework. However, the shape of the curvigram is very sensitive to the measurement uncertainty: for larger uncertainties fewer peaks will be created, and the location of said peaks also shifts. The

measurement uncertainty is usually estimated by the fieldworker so it is an input parameter of this method, that needs to be taken into account.

The quantitative approach implemented in this paper mirror the curvigram method. Using R, as above, a bespoke algorithm that constructs the curvigram from a list of measurements and associated uncertainties was coded. The algorithm then identifies its maxima by looking at the first and second derivatives of the curvigram and outputs the values of the highest peak, as we are only considering a single intentional target. For simplification, in all simulations the same value of measurement uncertainty was used for all measurements, even though that value was allowed to vary.

The Maximum Likelihood Method

There is another approach that, despite being considered the most natural approach to inference from a series of measurements (Taylor 1997: 93-120), and as far as this author is aware, has been very rarely applied by archaeoastronomers: to get the maximum likelihood estimate of the dataset. The *method of maximum likelihood* (ML) simply states that, given a number of empirical observations, the best estimate for their limiting distribution is that for which the empirical observations are most likely to occur (e.g. Edwards 1992: 70-102). This is a top-bottom approach to the problem of inference, since it is answering the question of what is the limiting distribution that maximises the likelihood of observing the set of measurements. From a theoretical point of view this is a very robust approach and, since its inception with Fisher (1922), the ML method has become the most widely used statistical approach to the estimation of parameters, given a set of empirical data, and is also part of the underlying framework of Bayesian statistics (e.g. Edwards 1992, 43-69).

Sophisticated as it may be, when applied to the single target scenario explored in this paper, this approach is very simple, as the maximum likelihood estimate is simply the mean of the dataset, in the case where all measurements have the same associated uncertainty. In the case of differing uncertainties, a weighted mean should be taken instead. From this perspective, Hoare and Sweet (2000) did use the mean for their inference of the orientation of early medieval churches in England, although whether they were aware of its inferential power or thought of it as merely descriptive is unclear. Furthermore, a clear advantage of this approach is that the inferential error can be derived from a simple equation: the inferential error is known as the *standard deviation of the mean* (Taylor 1997: 102) and, to calculate it, one needs only the total deviation in the data and the sample size (see Supporting Information for more on this and how it relates to precision).

In contrast with the Curvigram method, the ML approach is less prone to visualization, except that, after calculating the ML-inferred values, one can construct the inferred limiting distribution (figure 5). This figure, which used the same data as figures 3 and 4, shows that the ML-inferred limiting distribution is a close match to the actual one.

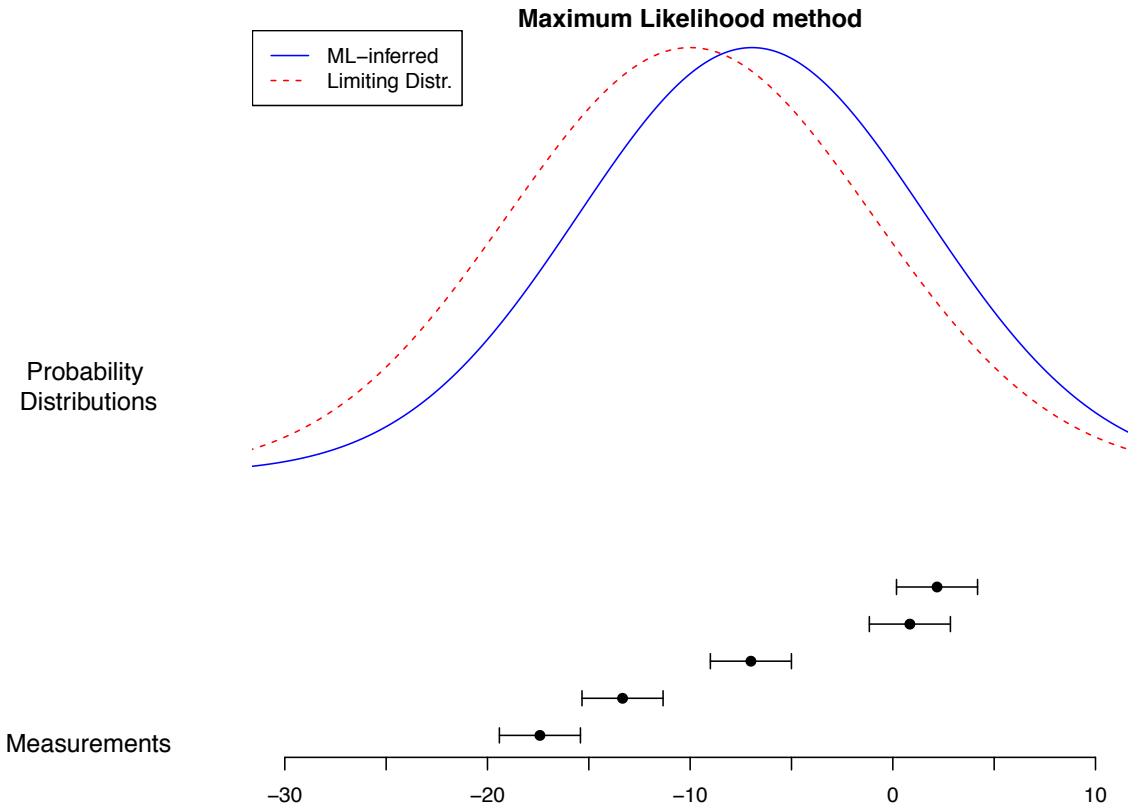

Figure 5 – Limiting Distribution inferred using the principle of Maximum Likelihood (solid blue curve at top) based on the same underlying data as those of figure 4 (dots at bottom). Compare the ML-inferred limiting distribution with the actual limiting distribution from which the measurements were random samples of (red dashed curve).

The implementation of the ML methods used in the analysis of simulation results were quite simple: as we are focused on a single target scenario, the mean of the (simulated) orientations was calculated.

Estimating the accuracy and precision of the methods

The difference between an inferred value and the true value is known as the *disturbance* or *error*. The lower this difference the better the method is at inferring the target of the alignment for that specific dataset, and so the more accurate it is. Due to the stochastic nature of the model this value is going to be different for different iterations of the Monte Carlo sampling. Furthermore, when using the inference method on an empirical dataset one cannot know whether the estimate is correct or, if not, what is the error. All one can know is how likely the method is to be correct and this can be estimated by analysing the distribution of the errors (e.g. figure 6 below) that results from the Monte Carlo simulations.

When the distribution of errors is obtained it is quite straightforward to calculate the expected accuracy and precision. The accuracy is given by the average of the errors. In the examples given in figure 6 both distributions peak at or around 0°, and indeed their averages are very close to that value, indicating that both methods are quite accurate. However, their precision is significantly different. Precision is associated with the width of the distribution, just as it is with the width of the cluster of arrows shot at the target in figure 3. It is conventional in statistics to consider a standard measure of the width of a bell-shaped

distribution, known as a standard deviation. The area covered by a single standard deviation corresponds to 68.2% of the entire distribution, whereas the area covered by two standard deviations comprises 95.4% of the distribution. These are often called the confidence intervals, as we can say that we have 95.4% confidence that a value drawn at random will fall in the area encompassed by two standard deviations, or 68.2% confident that it will fall in the area given by a single standard deviation. A confidence interval of 95%, corresponding to 1.96 standard deviations, is the most commonly used in statistical applications (e.g. Zar 1984, 43-45) and, therefore, we will use that value to define precision in this paper.

To illustrate this, figure 6 shows the error distribution for each of the methods being considered, for a scenario of 10,000 simulations, using parameter values taken from the first case study. The two histograms clearly show the ML method to be far more precise, with a much narrower distribution. In fact, this method infers an orientation that is within two degrees of the correct value 99.9% of the time. But we have defined precision at the 95% confidence interval so we need to look for that range (grey band). The 95% confidence interval falls at 1.26°, therefore giving us an inferential precision of 1.26°. On the other hand, at the same confidence level, the SPD method has a considerably different precision of 5.17°. In other words, under this particular scenario, we can be 95% confident that the target inferred by the ML is within 1.26° of the true value, whereas we can only say that the SPD-inferred target is within 5.17° of the correct value.

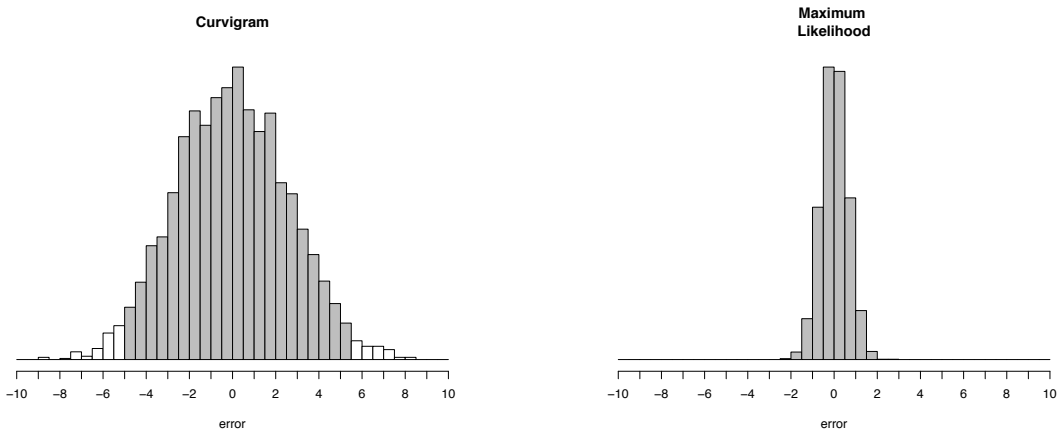

Figure 6 – Distributions of the error, that is the difference between inferred and true value, over ten thousand simulations of similar characteristics to those of figures 6-8, using the Curvigram (left) and Maximum Likelihood (right) methods. The grey colour highlights the 95% confidence interval of the methods.

In general, the method's accuracy and precision will change for different parameter values and, to explore this, a range of realistic values for the parameters were varied. To help with visualization, the accuracy and precision of each method, at 95% confidence, are presented below as coloured graphs where each axis corresponds to one parameter, and the colour scale encodes the accuracy or precision (depending on the case), from green (lower than 1°) to red (higher than 10°).

Results and Discussion

The accuracy obtained for both the Curvigram and the Maximum Likelihood methods were quite high, with little variation with changing parameter values. As figures similar to those shown for precision below (figures 8 and 9) would be all green, and therefore provide little information, boxplots for the accuracy of each method are shown instead (figure 7). As this makes clear, both methods are quite accurate: the Curvigram method is accurate to within one degree 96.77% of the time, with higher values found only for large deviations and small sample sizes; whereas the ML method is accurate to the same level 99.97%, displaying a slight advantage.

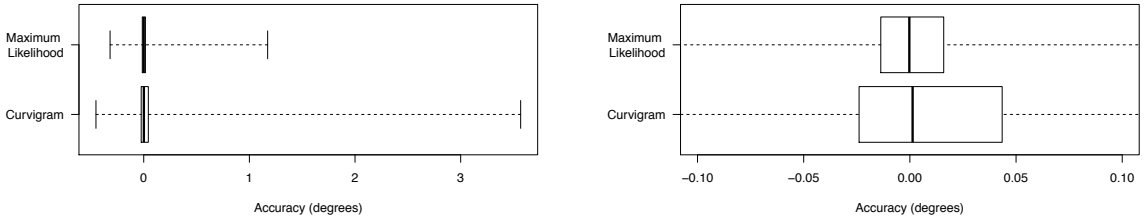

Figure 7 – Boxplots of the distribution of accuracy for the Curvigram and ML methods for all parameter value combinations. The left figure shows the full distribution, whereas the one on the right zooms in around 0° of accuracy.

Despite this impressive result, accuracy, on its own, tells us very little. An accurate method is one that, on average, gives us the right result. But, when applying it to a empirical dataset, one can never know if one is working on 'average' conditions, or whether the result will lie elsewhere on the error distributions of figure 6. Being accurate is definitely a good thing – an inaccurate method should be discarded *a priori* – but its precision also needs to be checked.

The precision of each method, however, varied widely. Figure 8 represents several combinations of parameters, two by two, for the Curvigram method, with the colours indicating the precision found for that combination of parameter values (see figure SF1 in Supporting Information for different combinations). As discussed above, the Curvigram method is sensitive to the estimate of measurement uncertainty (a key parameter in this approach), whereas the ML approach doesn't require it (so long as all measurements have the same uncertainty). For this reason, figure 9 is simpler, showing the only two parameters that are material to the ML approach.

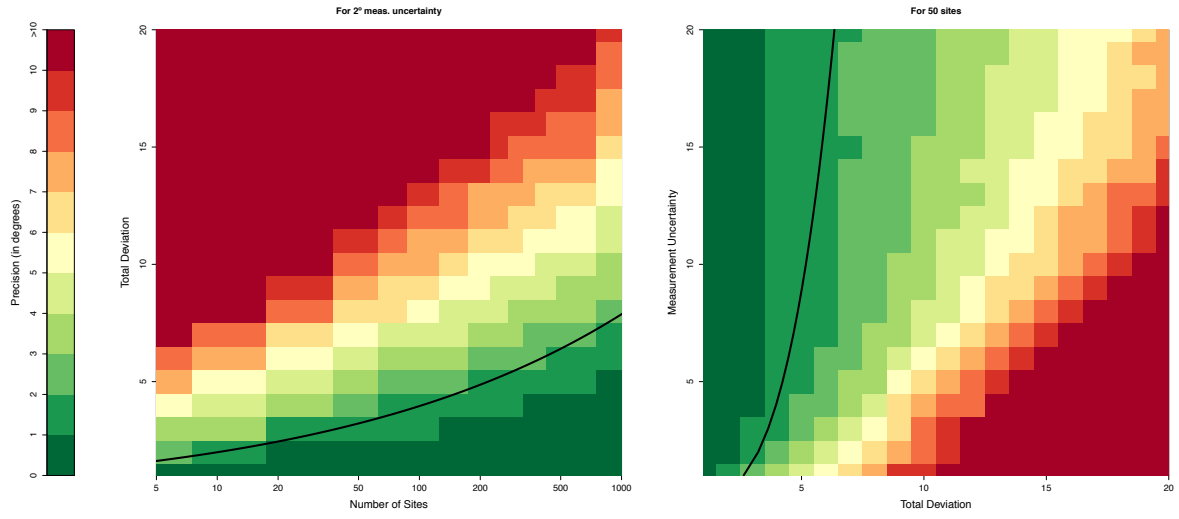

Figure 8 – Inferential precision for the Curvigram method for the single target scenario and for varying levels of measurement uncertainty, deviation from target and number of surveyed sites. The left figure is directly comparable to that for the ML approach (figure 9). The black lines are a fit to 2° precision (see text).

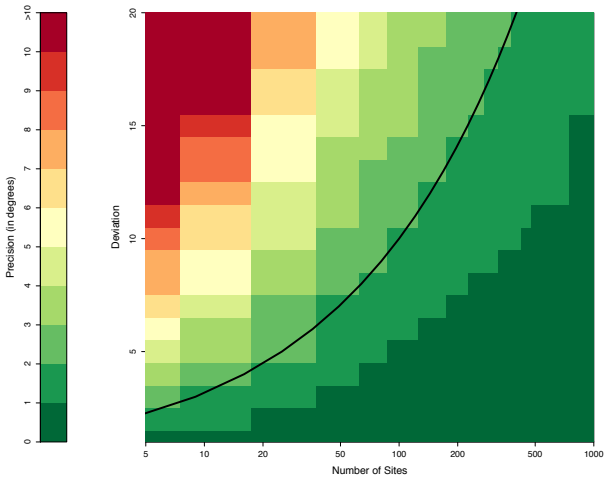

Figure 9 – Inferential precision for the ML approach for the single target scenario and for varying levels of deviation from target and number of surveyed sites. The figure is directly comparable to those for the Curvigram approach (figure 8, left). The black line corresponds to the theoretical expectation for 2° precision (see text).

The precision of the Curvigram method varies widely for different parameter combinations. A first observation is that, as expected, for a given value of deviation and measurement uncertainty, increasing the number of sites increases the precision (figure 8, left). One can then talk of a minimum number of sites required to achieve a certain level of precision. The black line on figure 8 is a power-law fit to the 2° precision band and shows the minimum number of sites for a given value of deviation and a measurement uncertainty of 2° (see details of its derivation in section 3.1 of the Supporting Information).

The same reasoning applies to the ML approach (figure 9). However, it is quite clear that this method is much more precise than the previous one. A direct qualitative comparison of the two figures reveals the ML method to have 2° precision or better over a wider range of parameters. Another way to look at it, is to note that, for a given total deviation, the ML approach requires a smaller number of sites to reach the same level of precision as the SPD approach. This is true irrespective of the measurement uncertainty used, as figure 10 shows.

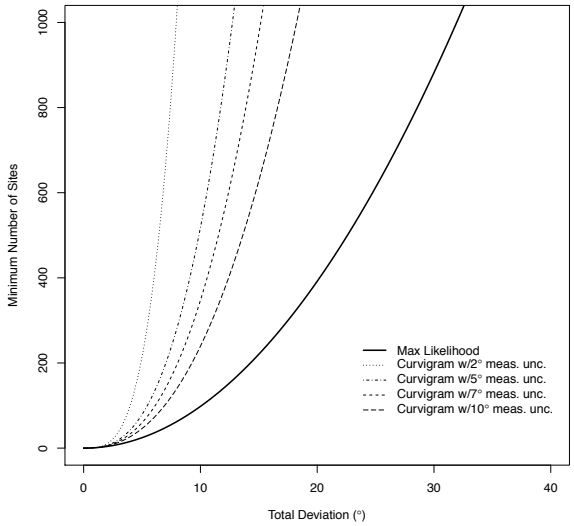

Figure 10 – Minimum number of sites to ensure an inferential uncertainty not greater than 2° for the ML method (thick solid line), and for the Curvigram methods with varying measurement uncertainties (dashed lines).

Figures 9 and 10 highlight another important point: that one's estimate of measurement uncertainty, used in the creation of the Curvigram, cannot be lower than the total deviation present in the data – in fact it should be much higher (see figure SF3 and its discussion in Supporting Information). When the measurement uncertainty is too low, the precision of the Curvigram inferences decays considerably; in other words, the inferred target is most likely wrong by several degrees. Figure 10 shows that, by increasing the measurement uncertainty the precision of the Curvigram gets closer to that of the ML method, further adding to the benefits of using the ML approach for the single target scenario.

These observations are important in themselves, but much more useful for future scholarship would be to have estimates of precision or, conversely, to know what is the minimum number of sites required to ensure that the precision is high. The data output by the Monte Carlo can be used to fit a multivariate equation that estimates the uncertainty from the sample size, total deviation and measurement uncertainty. However, because this relationship is non-linear and quite complex, it is only reproduced in the Supporting Information. A simpler equation to derive is that which gives the number of sites needed to ensure the inferential precision is small, say 2°. Its full derivation is also given in the Supporting Information, but the obtained equation is:

$$\text{(Curvigram)} \qquad N \geq dev^{\frac{1}{[0.247799+0.074932*\log(mes)]}} \qquad \text{(E1)}$$

where N is the number of sites, dev is the total deviation from the target and $log(mes)$ is the natural logarithm of the measurement uncertainty.

Turning to the ML method, one doesn't need to fit an equation to the data, since the obtained precision is perfectly predicted by the equation for the standard deviation of the mean (see

derivation in Supporting Information). In this case, the equivalent to equation E1 for the ML method is:

(ML) $$N \geq 0.98\, dev^2 \qquad (E2)$$

Therefore, to ensure that one's method is inferring the true target within 2°, one needs to have a number of sites that is equal to or larger than the values calculated using equations E1 or E2, depending on which method is chosen. In the Supporting Information, these two equations are compared to show that, under all realistic circumstances, in order to reach high precision the ML method will always require a smaller sample size than the Curvigram method.

In concluding the analysis of the simulations done under the single target scenario we can highlight two points. Firstly, the *precision of the Curvigram method is severely affected by underestimating one's measurement uncertainty*. And secondly, although both methods considered are sufficiently accurate, *the ML method is, under all circumstances, far more precise than the Curvigram method*. The provided equations (E1 and E2 above, together with SE1 and SE8) will allow scholars to estimate the precision of targets inferred from empirical datasets and report them in publications.

Conclusion

In this methodological paper, simulations were used to quantify the accuracy and precision of two methods of statistical inference that can be applied to the study of structural orientations in search of celestial and topographic alignments. The Curvigram, or Summed Probability Distribution, method is the most widely applied quantitative method in archaeoastronomy, despite it belonging to the realms of descriptive statistics, and hence more appropriate for exploratory analyses than for statistical inference. The analyses conducted above reveal this method to be highly imprecise under certain conditions, particularly when applied to datasets with small sample sizes, high total deviation (i.e. "noisy" data) and when the measurement uncertainty is underestimated (i.e. when it is well below the total deviation present in the data).

This paper has also introduced a method that, despite being of wide currency in other disciplines, had never been applied to the study of structural orientations: the Maximum Likelihood method. This method is far more reliable than the Curvigram method at inferring the intentional target of an alignment, demonstrating why it has become a cornerstone of modern statistics. Furthermore, equations that will allow other scholars to calculate or estimate the level of inferential precision of each method when applied to a specific empirical dataset have been provided: in the case of the Curvigram method the equation was fitted to the simulated data, whereas in the case of the ML method it has an algebraic derivation from a theoretical foundation.

However, the research here presented is but a stepping stone in the direction of a more robust, likelihood-based, statistical inference framework for skyscape research. Much more work is needed before it is useable in most scholarly applications. This paper has necessarily constrained itself to the simplest possible situation involving a single fixed-declination target. Nevertheless, the results apply equally to a fixed-azimuth target, and hence to both landscape

and skyscape analyses. The question of how the accuracy and precision of these methods changes for non-normal measurement distributions, for non-normal target distributions and for scenarios with multiple targets remains open.

The former, in particular, would be extremely important to address in the near future, but the large uncertainties observed for the single target scenario will likely reflect similarly large uncertainties in the presence of multiple targets. This raises the problem of resolution: what is the minimum target separation so that a statistical method can resolve between them? A second important issue relates to the choice of number of targets: the Curvigram method usually displays more than one peak, whereas a Gaussian Mixture Model – the ML approach equivalent for multiple targets (cf. McLachlan and Peel 2000) – requires one to choose *a priori* how many targets are present. In the absence of independent cultural data that can be used to subset the data, one needs a robust selection approach that should be properly tested against controlled simulated data, using a methodology not dissimilar to that of this paper. These points will be the focus of future papers.

Only when these issues are tackled, and a full framework developed and tested, will we be in a solid position to analyse case studies, re-assess past statistical analyses, and potentially identify datasets that might have been incorrectly inferred to be aligned with certain celestial or topographical targets. Until then, histograms and curvigrams will continue to provide useful descriptive information about collected orientation data. But the presented results begin to raise serious concerns over interpretations of empirical datasets based on such descriptive approaches, especially those that relied solely on quantitative analysis, as was the case for the 'green' school of archaeoastronomers. As mentioned, there is still considerable work needed to understand the (in)precision of the Curvigram method in more realistic situations, and how well a likelihood-based approach will fare under such circumstances, but this first analysis seems to suggest that a paradigm shift within quantitative cultural astronomy would be of significant benefit.

Acknowledgments
The author would like to acknowledge Liz Henty, who carefully read this manuscript and provided feedback, as well as the three anonymous reviewers whose invaluable feedback greatly improved this paper. Any lingering mistakes are my own.

References

Aveni, Anthony F. (ed), 1982. *Archaeoastronomy in the New World: American Primitive Astronomy*. Cambridge: Cambridge University Press.

Barcelo, Juan A. and Igor Bogdanovic (eds), 2015. *Mathematics and Archaeology*. New York: CRC Press.

Belmonte Avilés, Juan Antonio, 2006. De la arqueoastronomía a la astronomía cultural. *Boletín de la SEA* 15: 23-40.

Published in *Journal of Skyscape Archaeology* 3(1): 93-111. DOI: 10.1558/jsa.31958

Bevan, Andrew and Mark Lake (eds), 2013. *Computational Approaches to Archaeological Spaces*. Walnut Creek: Left Coast Press.

Clapham, Christopher and James Nicholson, 2014. T*he Concise Oxford Dictionary of Mathematics 5th edition*. Oxford: Oxford University Press.

Cummings, Vicki, Andrew Jones and Aaron Watson, 2002. Divided Places: Phenomenology and Asymmetry in the Monuments of the Black Mountains, Southeast Wales. *Cambridge Archaeological Journal* 12(1): 57-70. Doi: 10.1017/S0959774302000033

Edwards, A.W.F., 1992. *Likelihood. Expanded Edition*. Baltimore and London: The Johns Hopkins University Press.

Fishman, George S., 1995. *Monte Carlo: Concepts, Algorithms, and Applications*. New York: Springer.

Fisher, Raymond A., 1922. On the Mathematical Foundations of Theoretical Statistics. *Philosophical Transactions of the Royal Society* A 222: 309-368.

Freeman, P.R. and W. Elmore, 1979. A test for the significance of astronomical alignments. *Archaeoastronomy* 1 (Supplement to the *Journal for the History for Astronomy* 10): S86–S96.

González-García, A. César and Ivan Sprajc, 2016. Astronomical significance of architectural orientations in the Maya Lowlands: A statistical approach. *Journal of Archaeological Science: Reports* 9: 191-202. DOI: 10.1016/j.jasrep.2016.07.020

Heggie, Douglas C., 1982. *Archaeoastronomy in the Old World*. Cambridge: Cambridge University Press.

Henty, Liz, 2014. “The Archaeoastronomy of Tomnaverie Recumbent Stone Circle: A Comparison of Methodologies”. *Papers from the Institute of Archaeology* 24(1), Art. 15 (online edition). doi: 10.5334/pia.464

Hoare, Peter G. and Caroline S. Sweet, 2000. “The orientation of early medieval churches in England”. *Journal of Historical Geography* 26(2): 162-173. Doi: 10.1006/jhge.2000.0210

Hoskin, Michael, 2001. *Tombs, Temples and Their Orientations: A New Perspective on Mediterranean Prehistory*. Bognor Regis: Ocarina Books.

Hoskin, Michael, 2002. Studies in Iberian Archaeoastronomy: (9) An Overview. *Archaeoastronomy* 27 (supplement to *Journal for the History of Astronomy* 33): S75-S82.

Lake, Mark W., 2014. Trends in Archaeological Simulation. *Journal of Archaeological Method and Theory* 21(2): 258-287. Doi: 10.1007/s10816-013-9188-1

Magli, Giulio, 2016. *Archaeoastronomy: Introduction to the Science of Stars and Stones*. New York and London: Springer International Publishing.

McLachlan, Geoffrey J. and David Peel, 2000. *Finite Mixture Models*. New Jersey: John Wiley & Sons.

Pimenta, Fernando , Luis Tirapicos and Andrew Smith, 2009. A Bayesian Approach to the Orientations of Central Alentejo Megalithic Enclosures. *Archaeoastronomy* XXII: 1-20.

Pimenta, Fernando, N. Ribeiro, A. Smith, A. Joaquinito, S. Pereira and L. Tirapicos, 2015. "Open air rock art between Alva and Ceira rivers: a voyage through mining, trading, transhumance routes and the orientation in the landscape". In *SEAC 2011 Stars and Stones: Voyages in Archaeoastronomy and Cultural Astronomy*, edited by F. Pimenta, N. Ribeiro, F. Silva, N. Campion, A. Joaquinito and L. Tirapicos, 220-230. Oxford: British Archaeological Reports (International Series 2720).

Prendergast, F., 2011. *Linked landscapes: spatial, archaeoastronomical and social network analysis of the Irish Passage Tomb tradition*. Ph.D. thesis, University College Dublin

R v.3.3.2 https://www.R-project.org Accessed October 2016

Ruggles, Clive L.N., 1981. A critical examination of the megalithic lunar observatories. In *Astronomy and society in Britain during the period 4000–1500 BC*, edited by Clive L. N. Ruggles and Alistair W.R. Whittle, 153-209. BAR British Series 88. Oxford: British Archaeological Reports.

Ruggles, Clive L.N., 1999. *Astronomy in prehistoric Britain and Ireland*. New Haven: Yale University Press.

Ruggles Clive L.N. (ed), 2015. *Handbook of Archaeoastronomy and Ethnoastronomy*. New York and London: Springer.

Shennan, Stephen, Sean S. Downey, Adrian Timpson, Kevan Edinborough, Sue Colledge, Tim Kerig, Katie Manning and Mark G. Thomas, 2013. Regional population collapse followed initial agriculture booms in mid-Holocene Europe. *Nature Communications* 4:2486. Doi: 10.1038/ncomms3486

Silva, Fabio, 2015. "The View from Within: a 'Time-Space-Action' Approach to Megalithism in Central Portugal". In *Skyscapes: The Role and Importance of the Sky in Archaeology*, edited by Fabio Silva and Nicholas Campion, 120-139. Oxford: Oxbow Books.

Silva, Fabio and Fernando Pimenta, 2012. The Crossover of the Sun and the Moon. *Journal for the History of Astronomy* 43(2), 191-208. doi: 10.1177/002182861204300204

Silva, Fabio, CJ Stevens, A Weisskopf, C Castillo, L Qin, A Bevan and Dorian Q Fuller, 2015. Modelling the Geographical Origin of Rice Cultivation in Asia Using the Rice Archaeological Database. *PLoS ONE* 10(9): e0137024. doi: 10.1371/journal.pone.0137024

Silverman, Bernard W.,1986. *Density Estimation for Statistics and Data Analysis*. London: Chapman & Hall/CRC.

Sims, Lionel, 2009. Entering, and returning from, the underworld: reconstituting Silbury Hill by combining a quantified landscape phenomenology with archaeoastronomy. *Journal of the Royal Anthropological Institute* 15(2): 386-408. Doi: 10.1111/j.1467-9655.2009.01559.x

Sprajc, Ivan, 2015. "Astronomical Correlates of Architecture and Landscape in Mesoamerica". In *Handbook of Archaeoastronomy and Ethnoastronomy*, edited by Clive L. N. Ruggles, 715-728. New York: Springer.

Stevens, Chris J. and Dorian Q Fuller, 2012. Did Neolithic farming fail? The case for a Bronze Age agricultural revolution in the British Isles. *Antiquity* 86: 707-722.

Taylor, John R., 1997. *An Introduction to Error Analysis: the study of uncertainties in physical measurements*. Sausalito, CA: University Science Books.

Thom, Alexander, 1955. *Megalithic sites in Britain*. Oxford University Press: Oxford.

Thom, Alexander, 1967. *Megalithic lunar observatories*. Oxford University Press: Oxford.

Tilley, Christopher, 1997. *A Phenomenology of Landscape: Places, Paths and Monuments*. London: Bloomsbury Academic.

Williams, A., 2012. The use of summed radiocarbon probability distributions in archaeology: a review of methods. *Journal of Archaeological Science* 39: 578-589.

Zahid, H. Jabran, Erick Robinson and Robert L. Kelly, 2015. Agriculture, population growth, and statistical analysis of the radiocarbon record. *Proceedings of the National Academy of Sciences* 113 (4): 931-935. Doi: 10.1073/pnas.1517650112

Zar, Jerrold H., 1984. *Biostatistical Analysis*. New Jersey: Prentice Hall International.