Breakage, bias and the archaeological surface record: Assessing the quantification problem in archaeological field survey

Waagen, J.

DOI
10.1111/arcm.12720

Publication date
2022

Document Version
Final published version

Published in
Archaeometry

Received: 2 September 2020 | Accepted: 15 September 2021

DOI: 10.1111/arcm.12720

ORIGINAL ARTICLE

Breakage, bias and the archaeological surface record: Assessing the quantification problem in archaeological field survey

Jitte Waagen

Amsterdam Centre for Ancient Studies and Archaeology (ACASA), University of Amsterdam, Amsterdam, The Netherlands

Correspondence
Jitte Waagen, University of Amsterdam, Amsterdam Centre for Ancient Studies and Archaeology (ACASA), Amsterdam, The Netherlands.
Email: j.waagen@uva.nl

Funding information
The Netherlands Organisation for Scientific Research (NWO)

Abstract
In the practice of archaeological field survey there is a manifest importance for densities, that is, an abundance of artifacts, often relying on simple counts of objects. However, a well-known issue is variable breakage of pottery that can cause biases in quantitative analysis. Although such issues are generally acknowledged, a direct assessment of breakage and the resulting biases is lacking in research. This paper explores the pros and cons of quantification methods of surface collections in terms of counts or weights, and demonstrates the importance of analysing weight and breakage as part of an integrative approach.

KEYWORDS
archaeological field survey, bias, breakage, fragmentation, point samples, pottery abundance, quantification, statistics, surface record

INTRODUCTION

In the practice of archaeological field survey, archaeological artefacts lying on the surface are mapped and collected to study them as material correlates of past human activity. Whereas exact strategies depend on theoretical and practical considerations, the aim of archaeological field survey is generally to arrive at a representative sample of the find distributions present in the area. Common is the so-called regional 'off-site' survey where the archaeological record is considered as a spatial continuum and its variable densities are recorded (Foley, 1981) by consistently line walking all, or a subsample of, the visible terrain. This usually entails a systematic collection of finds resulting in a sample collection that can then be compared with other archaeological and historical sources (Alcock et al., 1994; Banning, 2002; Barker, 1995; Bintliff

et al., 1999; Fentress, 2000; Stek & Waagen, forthcoming; Waagen, 2014; Witcher, 2006). Two of the most important parameters for analyses of these find distributions are the relative densities of collected fragments and assemblage composition, that is, relative densities of specific artefacts in proportion to each other: what finds do occur, how often, how abundant in relation to other finds and how much more than in other places? The typical products of an archaeological field survey, after a usually elaborate study of potential biases (cf. Given, 2004; van Leusen, 2002: 4.1–4.20), are therefore geographical information system (GIS) maps displaying relative densities, either absolute or proportional, of all or subsets of the collected finds. These can be total densities, or densities of specific ceramic ware classes (wares), finds that share a chronological phase, shape or functional interpretation, or statistical properties such as richness (variation in wares) and evenness (proportional distribution of different wares). Focal points are often called 'sites' or what can more technically be described as concentrations of surface finds likely related to human activity in the past at that location, often associated with buried archaeological strata. However variably designated, for example, abnormal densities above background, points of interest or concentrations, 'sites' are generally defined by shifts in proportional abundance of finds in combination with other variables, such as the specific composition of ware types, shapes or functional interpretation of surface samples, as well as an assessment of post-depositional processes and geomorphological context. So-called off-site finds are displayed as general trends in their distribution, again represented as fluctuating densities of find types in various compositions (Bevan & Conolly, 2004; Bintliff et al., 1999; Dunnell & Dancey, 1983; Foley, 1981; Gallant, 1986; Waagen, 2014). Such maps are either used for direct interpretation, or considered as heuristic tools in an integrated analysis together with other archaeological and historical sources such as excavation data, geomorphological studies, textual evidence, etc. Assessments of these maps are then characterized by spatial pattern analyses, which aim to link the variability in surface distributions recovered by the survey to human activity in the past.

Given the manifest importance of densities, that is, abundance, in the analyses of data generated by archaeological field survey, quantification of finds is pivotal in the interpretative process. Although there has been well-founded criticism on the intensive off-site survey method, the value of the resulting distribution maps and the potential of quantification of the data (cf. Blanton, 2001; Fentress, 2000; Terrenato, 2004), it is still recognized that counting and quantifying in this field are vital for integrative problem-oriented research (Fentress, 2000: 50). However, the basic quantification in archaeological field survey often relies on simple counts of objects, which can be problematic. In most arable environments the sheer majority of the collected artefacts consists of broken ceramic sherds, in various degrees of physical deterioration. Various natural and mechanical processes can affect the state of ceramic finds in the subsoil and on the surface (Ammerman, 1985; Barton et al., 1999; Taylor, 2000; Terrenato, 2004; Winther-Jacobsen, 2010), one of its properties being fragmentation, also known as breakage or brokenness (Orton et al., 2013: 166–181). Although fragmentation issues are known, and can even reflect useful patterns, variability is introduced by pre-, peri- and post-depositional processes, for example, agricultural activities such as ploughing or levelling. Given this potential for variable states of breakage, relative densities may be affected. Observation of ploughed surfaces suggests higher fragmentation rates of sherds than for unploughed surfaces, attested through observation (Dunnell & Simek, 1995: 308) and by statistical size comparison between the two (Palumbo, 2015: 86). Variability in ware-dependent fragmentation rates results in similar biases, that is, large pots of wares susceptible to high degrees of fragmentation will likely be overrepresented (Orton et al., 2013: 169; Strack, 2011: 24). Whereas such a bias may be overcome by comparing samples only based on a single ware and/or a limited range of vessel sizes, post-depositional biases are not so easy to avoid (Orton et al., 2013: 169). Although these issues are acknowledged by field survey practitioners, and various approaches have been developed (cf. Akkeraz & Collins-Elliot, 2017; Coccia & Mattingly, 1992; Tol, 2012;

Winther-Jacobsen, 2010), a direct assessment of these quantifiers and the degree to which they can result in biases is often lacking in research and the use of counts is still ubiquitous. This is tricky since such biases potentially affect research outcomes if not dealt with carefully. One may therefore argue that the quantitative fundament of the creation of distribution and site maps is a subject that calls for further examination. This paper explores the pros and cons of quantification methods of surface collections in terms of counts or weights.

COUNTING POTS AND POTTERY

Establishing a proxy for abundance, that is, a quantifier, is a topic that has been widely addressed in ceramic studies, and various techniques have been examined (Orton, 1975; Orton, 1993; Orton, 2000; Orton et al., 2013). Most authors stress that it is important for ceramic studies to move beyond mere counts and weights of sherds. There have been many different attempts to find a workable solution to the problem of quantifying ceramic assemblages, for both excavation as well as surface record assemblages (cf. Arcelin & Tuffreau-Libre, 1998; Dawson, 1971; Egloff, 1973; Fulford, 1973; Mateo Corredor & Molina Vidal, 2016; Orton, 1975; Poulain, 2013; Py, 1991; Strack, 2011; Verdan, 2011). An important branch of techniques that have been developed can be grouped under the term 'estimate of vessels represented' (EVREP), originally designed for use with closed contexts (Orton, 1993: 176). These methods attempt to circumvent the issue of breakage by applying a standardized approach to establish the number of whole vessels that the sherds are likely to represent. Some of the best known of these methods calculate the minimum number of individuals (MNI) and maximum number of individuals (MNA), or an average between the two. These rely on a count of feature sherds, that is, rims, handles, bases or those with other identifiable features, rather than the body sherds, of any given ware. Because their relative number is usually known, these can be counted and attributed to a minimum or a maximum number of pots from which they could stem. An alternative is the aggregate feature count (AFC) which is derived by adding sums of rims, handles and bases. This method assumes equal breakage rates for similar sized and types of vessel, and that all vessels present actually have rims, handles and bases (Strack, 2011: 24). The estimated vessel equivalent (EVE) is a non-representational quantifier (e.g., Egloff, 1973; Orton, 1982: 164–167; Orton, 1993: 172). Whereas a non-representational quantifier also strives to arrive at comparable quantities, this is attempted by constructing an abstract representation, that is, there is no claim of reference to actual pots represented. The EVE refers to quantification of a part of a pot for which its proportion to the whole vessel is known, alleviating the problem of variable breakage. Often-used is the rim-EVE, but other EVEs are possible, for example, a weight-EVE is possible in case of highly standardized weights per pot (Baumhoff & Heizer, 1959: 309; Raux, 1998: 12), a vessel surface-EVE has also been proposed (Byrd & Owens, 1997; Hulthén, 1974). Whereas the EVE is deemed useful as well as powerful (Orton et al., 2013: 173–174), understanding the abstract mathematical mechanics used to derive a metric that allows statistical analysis, the *pie slice*, may be challenging. With good reason, its creator aptly calls it 'a creature fit for a mathematical zoo' (Orton et al., 2013: 174), and probably the reason it is often not structurally used.

It is important to note that all these quantification solutions perform well in assemblages with a high level of completeness (percentage of the pot being present in the assemblage) and a low level of brokenness, and in most cases require an individual treatment of every single sherd. The obvious problem here for archaeological field survey is the generally worn physical condition of the collected finds, which are most often highly fragmented and incomplete. Feature sherds typically make up a small size of the sample; in the ceramic body collected by the Tappino Area Archaeological Project (TAAP), Molise (central–southern Italy), the percentage of such sherds hovers around 10% (Stek & Waagen, forthcoming). This renders approaches

based on feature sherds not very easily applicable, if not impossible, at least for Mediterranean archaeological field survey assemblages (Winther-Jacobsen, 2010: 50). To be able to account for all finds, and have sizeable and well-distributed samples, for many research purposes it is imperative to work with the large numbers of non-feature sherds (Strack, 2011: 24). Furthermore, the sheer size of the collected samples often inhibits individual assessment of the objects. Using a broad range of quantifiers side by side such as the proposed standard by the Seville protocol (Adroher Auroux et al., 2016) is nigh impossible to implement. Although in Greece individual assessment (e.g., Bintliff et al., 2017; Krijnen et al., Accepted/In press) seems to be more often part of the research tradition than in other Mediterranean areas, the above problems make weighting and counting the batches of sherds per ware or feature type the modus operandi.

A review of the arguments

In the scope of archaeological field survey where the bulk of the finds are badly preserved, the question is to what degree is counting or weighing sherds the least biased quantity estimator. The rationale for the comparison is that weight is not affected by breakage and therefore does not suffer from that bias when comparing based on ware types (Orton et al., 2013: 169). If breakage is constant, a choice for using counts can well be defended for archaeological field survey. It is a straightforward measure that relates to what is picked up in the field and connected to common thresholds for, for example, site identification. Using weight, on the other hand, introduces difficulties on its own, which will be addressed further below.

Previous research indicates preference of weight above counts (e.g., Carrete et al., 1995; Millet, 1991; Millett, 2000; Orton & Tyers, 1993). This is empirically demonstrated by strong correlations between various measures (e.g., numbers, weights, rim counts), so whichever measure was chosen, the trends in the abundancy estimates remain similar. Therefore, the choice of a good working parameter can be purely based on ease of implementation; weights of sample batches are by far the fastest to record (Millet, 1991). However, details of the tests are omitted, and they are likely not representative of all ware types and contexts, for example, based on wheel-made pottery from a single courtyard (Millett, 1979), limiting the degree to which the results can be generalized. Yet others look at the general correlation between the relative abundance of wares per period as expressed in counts or weights and draw similar conclusions (Slane, 2003: 325–326). Although they indeed attest such a correlation, they do not specify statistical significance, but more importantly, only very general trends are tested. The issue for the current problem is that taking batches of finds and correlating numbers and weights invariably introduces a smoothing effect; working with total or average weights, variation on an individual sherd basis is masked.

Some studies suggest that breakage rates do show averaging trends. The concept lies at the basis of the modulus of rupture (MR) (Molina Vidal, 1997; Mateo Corredor & Molina Vidal, 2016), which is an estimate for the average size of a sherd breaking off from a specific vessel. This modulus is then a corrector for breakage that can be applied to counteract differential breakage rates between wares and pot types. However, the assumption of a random breakage process, which is the basis for the MR coefficient, is unwarranted for survey archaeology due to variable post-depositional histories. Another argument for treating counts and numbers as roughly displaying the same trends is that whatever biases are there, they are likely to be consistent (Winther-Jacobsen, 2010: 49), an argument that, again, is invalid for archaeological field survey, where conditions of pottery preservation can be different for adjacent fields.

A final argument that prefers counts above weights is the perception that the statistical probability of picking up two sherds from a single pot is negligible, and thus one could use

counts as a sort of MNI (V. V. Stissi, personal communication). However, in cases of ploughed up deposits such as in a site context, this assertion appears hard to justify. Since buried contexts can show a high degree of completeness, it is statistically very possible that sherds originally did belong to the same pot, or to a larger surviving fragment. Admittedly, such connections between sherds, that is, 'sherd families' (Orton et al., 2013: 172), are difficult to detect due to the physical wear of break lines and, again, the usual abundance of finds. One usually does not invest in comparing all possible cases of fitting sherds, though where it has been, such cases have been signalled (Tol, 2012, 237–238). Also, as mentioned above, there is statistical evidence pointing towards breakage of surface finds due to ploughing (Palumbo, 2015: 86).

Weight issues

From the above, there is clearly no solid ground for assuming a constant relation between number and weight on a general level, or an uncomplicated argument to prefer counts over weights. The degree to which breakage is a factor influencing the quantitative basis of find densities is still largely one open for investigation. To recount, if breakage is very variable, weight is likely the less biased quantifier (Orton et al., 2013, 169), but weight brings its own specific limitations and potential biases.

The most obvious limitation is that large vessels, at least those with thick walls, will on average feature heavier sherds in comparison with small vessels, rendering in-sample comparisons of proportions skewed. Working with weight means that the individual ware classes must be considered in all comparisons (Orton et al., 2013: 169; Millett, 2000: 54). Solutions have been proposed, such as adjusted weight, surface correction and volume displacement (Hinton, 1977; Hulthén, 1974), but these are practically cumbersome. Average vessel weight, leading to a weight-EVE, would probably overcome most of those problems (Rice, 1987, 292); however, a lot of data would be necessary to make this work, and there is the issue of specific weight (Mateo Corredor & Molina Vidal, 2016), as mentioned below.

Similarly, a large range of vessel sizes within one ware will influence abundance estimates between samples, that is, if one sample contains sherds of small pots and another of large pots of a single ware class, quantity estimates based on weight will be skewed towards the latter. This is difficult to avoid and very dependent on the range of vessel sizes within one ware. Since wares are often based on fabrics with similar physical properties, such size ranges will not always be extreme, but certainly can be. Other potential biases may be caused by variability of weight within wares for similar vessels. After all, if sherd weight is not a robust proxy for sherd volume, there are similar issues as with using numbers as measures of abundance. Such may be caused by variable specific weight due to differences in the chemical composition of fabrics within a single ware, and because of processes such as overfiring and weathering. Since empirical research into the matter is lacking, albeit some explorations (Kinnunen, 2020), little is known about potential impact and solutions. A final issue concerning weights is loss of moisture due to evaporation. Ceramic finds can be assumed to be fully saturated with moisture at the time of collection, after which it will gradually reduce depending on the procedures of finds processing and storage. This has been demonstrated to lead to differences of 10–15% (Slane, 2003: 324), though this bias may be mitigated by timing the moment of measurement.

To conclude, where some of the limitations may be overcome, that is, by looking at wares separately, others, such as vessel size ranges and specific weight variability within wares, are more difficult to avoid and must be explicitly assessed. To examine the behaviour of count and weight proxies in more detail, this paper continues with an empirical statistical treatment of data from a case study.

A CASE IN POINT

In the context of the Tappino Area Archaeological Project (TAAP), coordinated by Dr Tesse D. Stek, University of Groningen and the Koninklijk Nederlands Insituut in Rome, a large-scale methodological test was designed for the investigation of a find complex interpreted as a large rural site, located on Colle S. Martino in the Tappino area (Fig. 1) (Stek, 2018; Stek & Waagen, forthcoming). The site was systematically examined by applying transect survey, where five people systematically collect all finds in a 2 m swathe spaced 10 m apart in units of roughly 50 x 50 m, that is, collecting a 20% sample by line walking (Pelgrom & Stek, 2010; Stek & Pelgrom, 2005; Waagen, 2014).

The relatively high densities (often more than 5/m^2), the size of the scatter (about 4 ha), the variable visibility and the spatial patterning of the artifacts made the case ideal for testing the efficiency of the so-called point sampling (PS) technique in relation to transect sampling (TS). PS is a very intensive sampling method in which a relatively small area is cleaned of its vegetation and very thoroughly examined for archaeological artifacts, in our case a circle with a 1 m diameter (Fig. 2).

Similar techniques, such as shovel-testing or test-pitting, have been experimented with quite a few times (Chadwick & Evans, 2000; Kintigh, 1988; Krakker et al., 1983; Lightfoot, 1989). Although these are comparable with PS in their spatial precision and aims for bias mitigation, PS is not an excavation: it is a surface examination and is relatively effort-efficient as opposed to digging pits. The method carries great potential to map find distributions with high precision avoiding common visibility and observation issues (Stek & Waagen, forthcoming). In order to test the PS method in comparison with regular TS, a grid for PS was laid out over the site, partly overlapping the already surveyed fields (Fig. 2). This enabled assessment in various visibility circumstances and densities because they were also placed in fields considered to be outside of the site scatter boundaries. The sample collection resulted in a data set consisting of 794 PS partly in non-visible terrain, partly overlapping 25 TS and a total of 9255 collected pottery fragments.

Clearly, the methodological study into PS required a good proxy for relative abundance for comparing differences in densities. Moreover, breakage is more prominent because of the

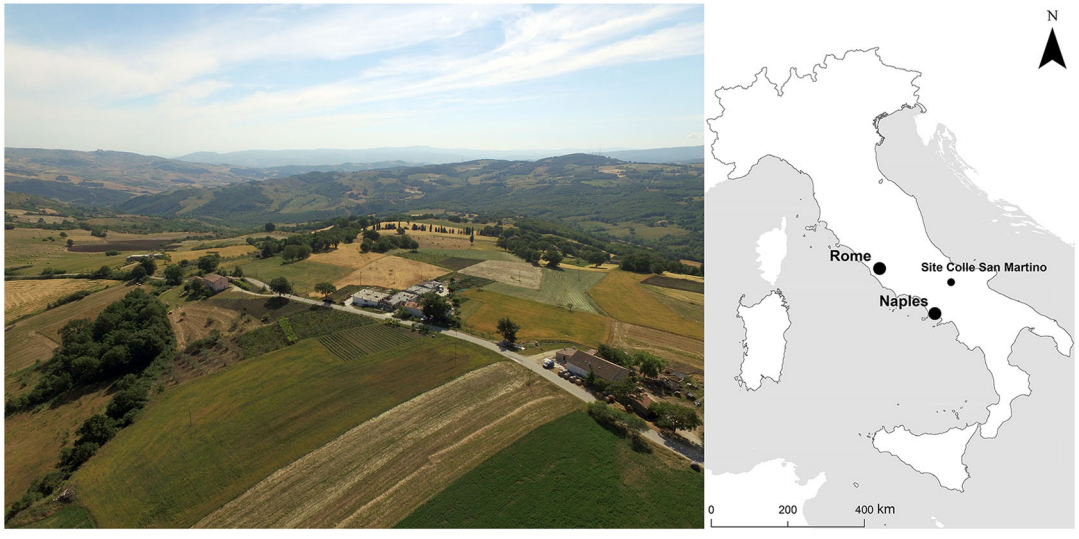


FIGURE 1 Site of Colle san Martino (photo: Tesse D. Stek) [Color figure can be viewed at wileyonlinelibrary.com]

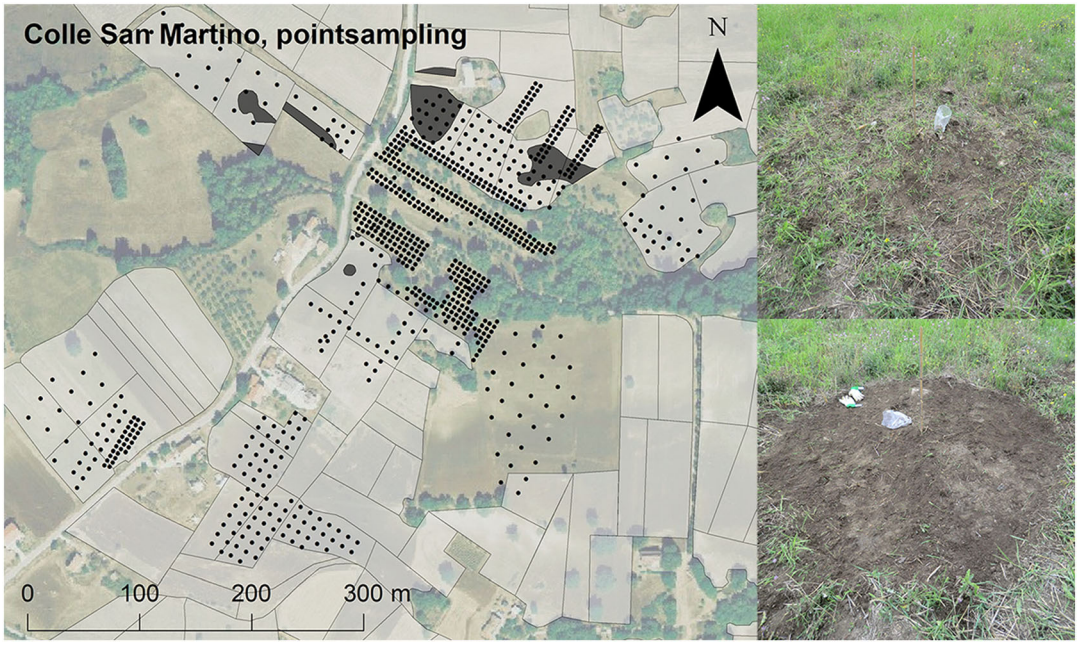


FIGURE 2 Point samples and transect samples at Colle san Martino: (left) distributions of point sampling (PS), black; transect sampling (TS), light grey; and site boundaries, dark grey; and (right) point samples before and after cleaning [Color figure can be viewed at wileyonlinelibrary.com]

increased intensity of the PS method. As opposed to TS, PS allow a very effective alternative for collection in a small area (Stek & Waagen, forthcoming). Consequently, the closer one looks to the ground surface, the smaller the pieces of ceramics that will be recovered. Surface sherds range from big to small, and more intensive sampling techniques, that is, time spent on collection, distance between the observer and the ground, etc., will target more effectively the smaller size ranges. Thus, one should be very careful when comparing counts of different sample types, as the smaller pieces may represent bits broken off larger pieces. Furthermore, the degree of breakage on a site such as this can be a source of evidence for the state of surface assemblages, and possible relation with buried deposits. Finally, it is known that when completeness is low, it is more likely for pots with a high level of fragmentation to be included in samples (Orton et al., 2013: 169). Especially, in case of the reduced sample area of PS, this could be of notable influence. All these factors render the study of breakage essential for the Colle S. Martino case, and thus formed an opportunity to engage with the problem.

Counts, weights and statistics

To examine breakage, the two most abundant wares, coarse wares (CW) and plain wares (PW) were examined to optimize statistical power. These wares constitute the main classes of wheel-made table and (light) utility ware in ancient Samnium. The bulk of these finds dates from the Archaic to Late Roman period, and in the TAAP data set most of them are found on sites with fourth- and third-centuries BCE black gloss (BG) pottery. The division between CW and PW is made on the presence or absence of larger inclusions in the clay matrix (Fig. 3). Although finer data would have been obtained by analysing finds of narrower chronological and morphological frames, such data unfortunately, except for a small subset of feature sherds,

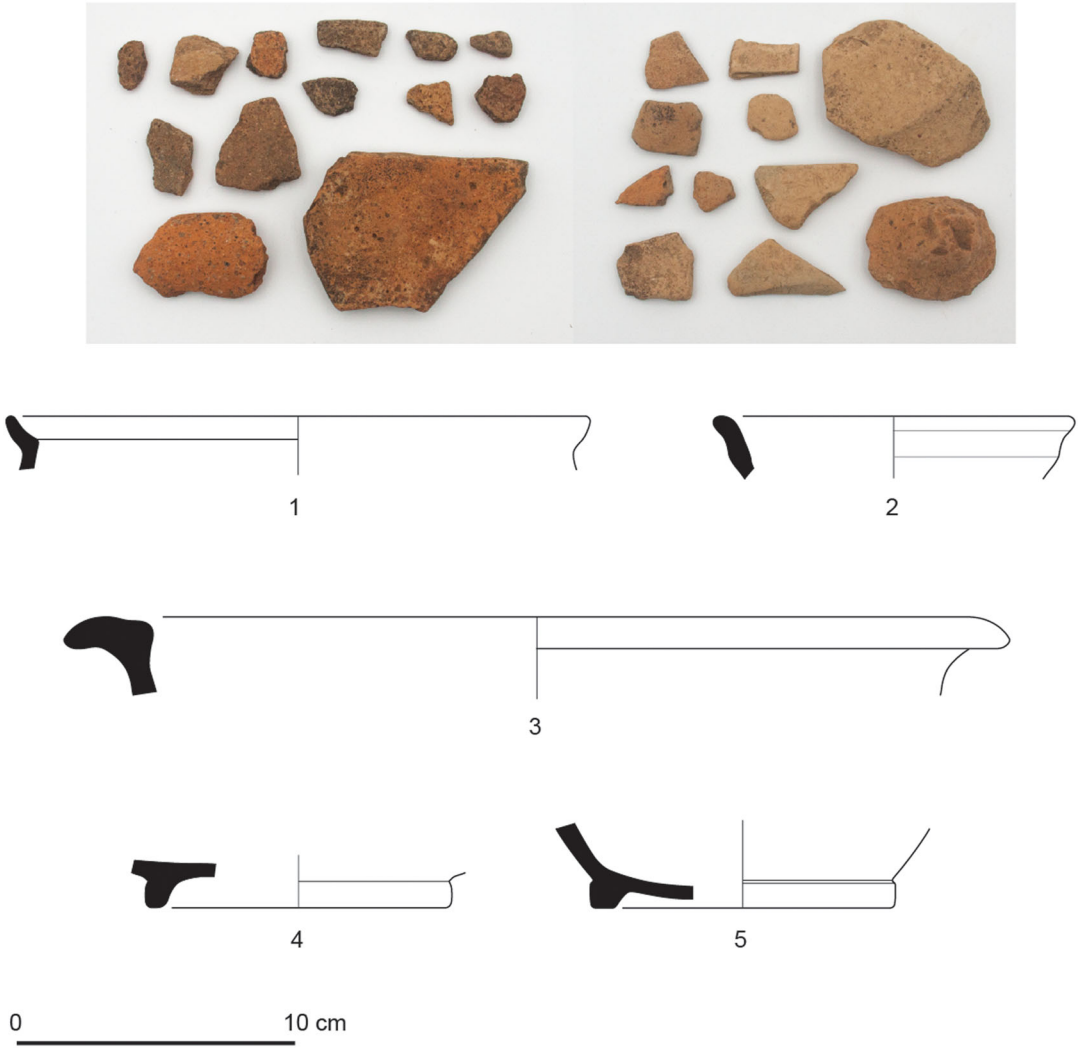


FIGURE 3 (top) Examples of coarse ware (left) and plain ware (right) pottery from the Tappino area clearly showing variable fragmentation; and (bottom) 1, coarse ware (CW) cooking pot; 2. CW bowl; 3. CW rim; 4, plain ware (PW) base; and 5, CW base [Color figure can be viewed at wileyonlinelibrary.com]

are simply not available. Apart from further identification of finds and groups being a question of effort expenditure, oftentimes the sherds themselves, being heavily corroded, do not allow for much further specification. Nevertheless, there are several indications that the larger share of these finds actually fit a narrower chronological frame, that is, about 60% of all CW and PW finds in the data set have been collected at 20 sites, of which at least 14 can be dated to the Hellenistic/Samnite phase based on BG pottery. Datable PW feature sherds follow a similar pattern. There are very few CW and PW finds that correlate with Iron Age or Archaic find distributions, and there is otherwise no evidence to suggest a very dissimilar chronological association of CW and PW off-site finds in the research area. It is a reasonable assumption that the majority of the finds under study are from the Hellenistic/Samnite period.

In the following assessment the weight issues mentioned above will be considered. Weight has been established in weighing batches of finds per ware per sample. The issue of variable

moisture is not very influential, as all finds have been weighted on the day of collection after washing. As they have come out of the field fully saturated, there is no great deal of moisture flux at this point. Although the effect of variable specific weight will be mitigated due to the far majority of the finds being of local or regional production, and all being wheel-made, this must be accepted as a potential source of some variability. When it comes to vessel size range per ware, these can create noise as well. The question is of course whether such noise will have a large effect. Very small or very large vessels of these wares may occur, but will likely be proportionally rare. Nevertheless, its effect must be estimated and therefore will be addressed further below.

Correlations

A first assessment is testing any correlation between number and weight to see whether or not the two show a similar trend. Taking all the CW from the full set of PS aggregated per TS, as well as the TS themselves, and comparing counts and weights with linear regression, there are strong and statistically significant correlations, respectively $r^2 = 0.95$, $p = 0.000$ and $r^2 = 0.96$, $p = 0.000$), against an $\alpha = 0.05$ significance level. For the PW in the PS/TS and the TS there is only a similar good match for the latter, that is, $r^2 = 0.92$, $p = 0.000$, but just a moderate correlation, still significant, for the PS/TS $r^2 = 0.66$, $p = 0.004$ (Fig. 4) against an $\alpha = 0.05$ significance level. Regression analysis assesses the degree to which variation in one variable behaves similarly to variation in another, so in the latter case the variation in counts of the sherds for 66% 'explains' the variation in weight of the sherds, which suggests that 34% of variation in weight is not related to differences in counts.

The 34% deviation is clearly attributable to two PS collections: one in TS 2322, a high-density area, and one in 2343, a low-density area. Removing them as 'outliers' would certainly result in a stronger correlation, but there is no evident reason to manipulate the data in this way. Being both collections of 13 PW sherds, but one with a total of 22 g and the other with a total of 117 g, one may wonder whether this is because of different vessel sizes or possibly variable because breakage due to post-depositional processes. Although the finds study did not clarify the issue, striking is a possible bowl fragment of 33 g in unit 2322, which appears rather

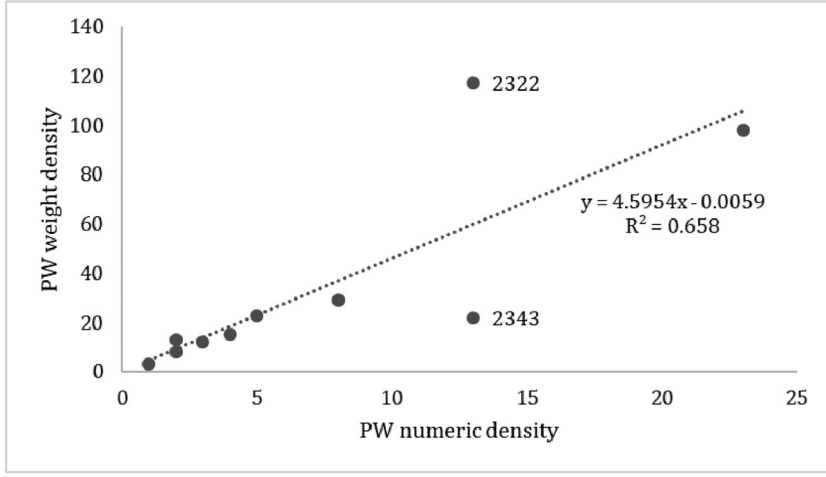


FIGURE 4 Regression analyses of plain ware (PW) pottery, point sampling (PS) and transect sampling (TS) combined

large. Both possibilities may potentially provide information on a possible difference between the nature of the assemblages, for example, site or off-site. In this way, studying breakage is provoking questions about the surface assemblages and may bring up new ideas about the site and its finds. More generally and importantly, however, it should be noted that the strong correlations all relate to the aggregations, that is, at a level of grouping of finds that in itself already smoothens variability by taking the mean sherd weight. On the level of individual PS, the same PW finds show an actually very weak correlation, and still statistically significant, that is, $r^2 = 0.14$, $p = 0.000$. Therefore, on a general level the correlation appears strong, but this correlation fails zooming in on individual samples, and variability becomes evident.

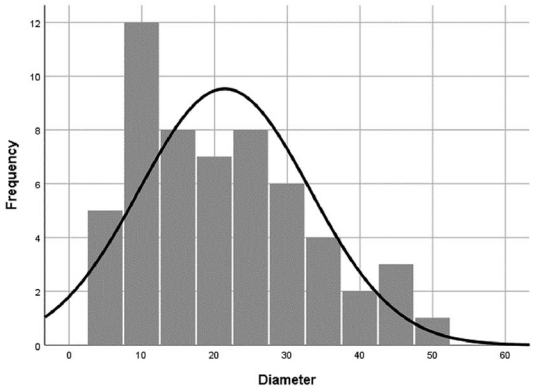
Variability

Whereas correlations allow assessment of trends, it is imperative to have a finer grip on the detail, for which common descriptive statistics can be applied, such as the coefficient of variation (CV). The CV is a descriptive metric that aims to provide information on the spread of a batch of measurements regardless of the mean (μ). Where the standard deviation (s) expresses an absolute measurement relative to μ, which is difficult to assess without μ itself, the CV gives the relative dispersion by dividing s by μ, arriving at a ratio, that is, the spread expressed as portion of the mean. This is best illustrated by an example: a 5 g s with $\mu = 10$ g is a lot of dispersion, whereas the 5 g s with $\mu = 100$ g is proportionally little dispersion; whereas s is the same, the CV of the former is 50% and that of the latter is 5%, expressing the difference of variation respective to μ. For assessing spread it is very useful to consider the CV alongside s and μ.

The CV for the PW finds from the PS is 42.72%, for the TS is 69.09%, and the average is 55.90%, which is considerable. To provide a comprehensive impression, this can be translated to the actual mean weight deviation. A total of 1 s from the mean, that is, notionally 66% of the finds, comprises a range of 2.55–6.35 g ($\mu = 4.45$ g, $s = 1.9$ g). In other words, for two-thirds of the finds, the heaviest pieces are 2.5× the weight of the least heavy sherds. For PW found in the TS, the variation soars to 69%, within 1 s ranging from 1.89 to 10.29 g ($\mu = 6.09$ g, $s = 4.2$ g). This points to a considerable weight variation in PW sherds, and potentially indicative of breakage effects.

A difficulty with assessing the potential effect of vessel size range is that there is limited information on that range, and that establishing it based on archaeological field survey finds is nigh impossible. A potential proxy is the distribution of diameter estimates taken from feature sherds. These show quite some variability, with $\mu = 21.4$ mm, $s = 11.7$ mm and CV = 54%, which is higher than the weight variability of PW sherds in PS and lower than that in TS (Fig. 5). In PS the average weight will be lower due to the collection of smaller pieces; however, the comparison with TS is more suggestive, since these are the regular samples featuring on distribution maps. However, the question remains to what degree vessel diameter estimates are a good indication of vessel size at any rate. Variability is likely introduced by differences between open and closed shapes, and the notion that the size of a vessel potentially increases faster than its rim diameters (Stissi, personal communication). Tests show that the only significant correlation between metrics such as sherd weight, vessel diameter and wall thickness is between diameter and wall thickness, $r = 0.47$ with $p = 0.000$. Whereas this provides some confidence to presume that vessel diameter is at least partly indicative of size, it is not the whole story.

Boiling this down to a single shape, that is, rim sherds of PW bowls, clearly there are variable weights because of larger and smaller fragments. These open shapes show some correlation in rim size and diameter, but the variability in weight does not always follow, for example, there are two rim sherds with a 0.6 cm wall thickness, one of 4 g and another of 10 g; two sherds with wall thickness 0.7 cm, and, respectively, 15 and 18 cm diameter, are, respectively, 5 and 15 g, not even close to a proportional increase in weight with diameter difference (Fig. 5).



FORM	FRAGMENT	WEIGHT	THICKNESS	DIAMETER
BOWL	r	4	0.6	12
BOWL	r	15	0.7	18
BOWL	r	3	0.4	12
BOWL	r	2	0.1	10
BOWL	r	5	0.7	15
BOWL	r	10	0.6	-
BOWL	r	4	0.5	16

FIGURE 5 Diameter distribution for plain ware (PW) feature sherds (left) and recorded metrics for point sampling (PS) bowl shapes (right)

Surely, this is a very small sample of a single shape, and it demonstrates the rather obvious: pots break in variably sized sherds independent of vessel size. Nevertheless, it eventually corroborates the notion that weight variability is partly related to breakage and not just to vessel size range. Although it may be hard to disentangle the actual cause for weight variability, all the presented analysis point towards an effect of breakage reflected in weight variability.

The CV weight difference between the PW samples in PS and TS show less variation in the PS. Since regression analysis showed that for PW in PS, an increase in count per PS/TS does only moderately cause an increase in weight, it is evident that PS collect the smaller pieces in the surface assemblage more effectively. The CV demonstrates here that the spread of weight per sherd for PS/TS is much smaller than that for the TS, corroborating the same pattern. The PS sample method gives up smaller finds because of a difference in intensity, obviously due to the ground being examined from a much shorter distance and for a longer time. These sherds might, at least for a part, be more fragmented parts of vessels, and thus counting numbers of sherds potentially overrepresents the total quantity of pottery collected. It is worth mentioning that this conclusion echoes the results from a seeding experiment, referring to the so-called *size effect* (Odell & Cowan, 1987).

The conclusion must then be that in general, there are good reasons to believe that weight variability is at least partly a result of breakage differences. This means that it is reasonable to consider that counts of sherds are a more biased estimation of densities than weight, and that in the case of PS, but more generally in case of more intensive sampling, biases can be expected to be stronger.

Effects on the spatial distributions

Whereas an elaborate treatment of the use of weights and counts for the study of Colle S. Martino is not the scope of this paper (Stek & Waagen, forthcoming), a few examples can be given to show that the biases created by using simple counts are not trivial. Plotting weights and counts per PS on the map, the variation is immediately visible (Fig. 6).

Evidently, where there is a general spatial correlation between the site boundaries and weight densities in TS 2320 and 2322, there are substantial deviations as well with the majority of the high-weight densities falling outside of the site boundaries (Fig. 6). Given the hypothesis that bigger sherds are often related to ploughed up deposits, and those finds will be

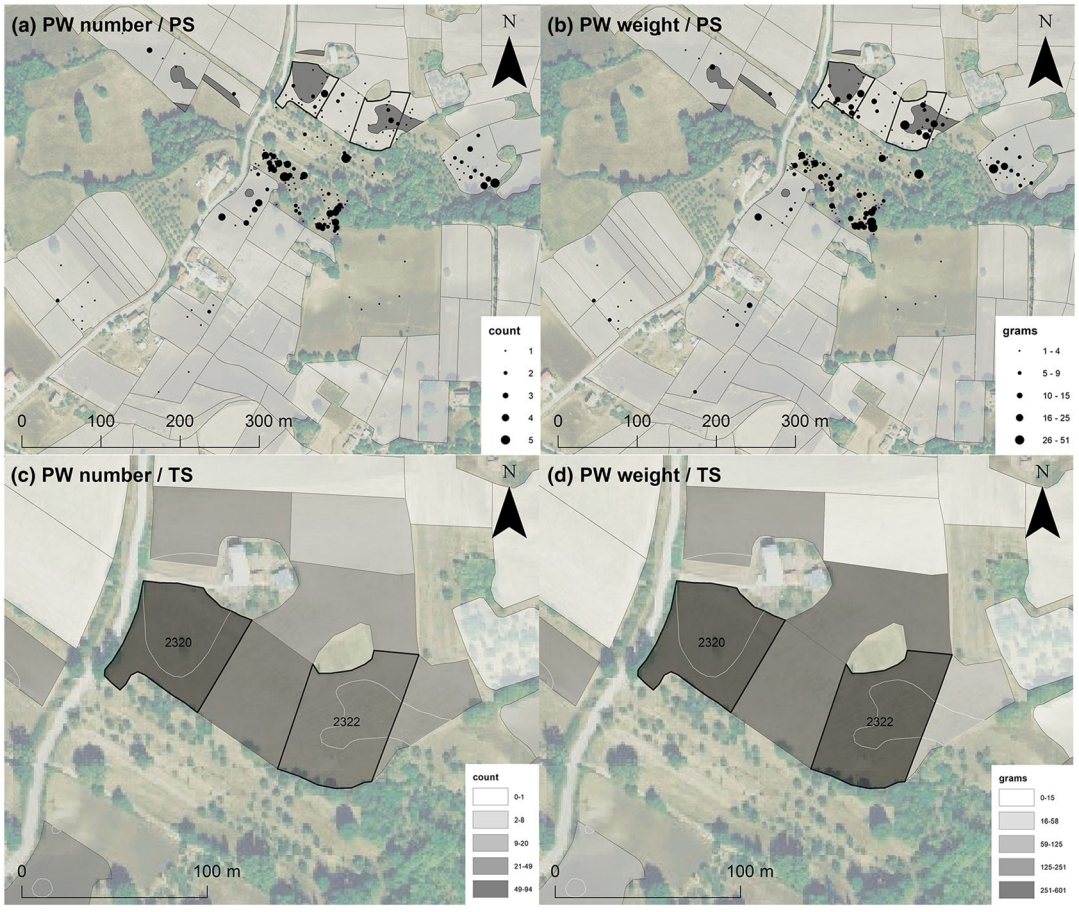


FIGURE 6 Numeric densities of plain ware (PW) pottery (A) versus weight densities of PW pottery (B): white lines indicate concentrations, and black lines TS 2320 and 2322, and PW densities, in number (C) and weight (D) [Color figure can be viewed at wileyonlinelibrary.com]

subsequently spread and further fragmented, this observation could point to site halos instead of the location of buried remains.

Using either numbers or weights of PW finds in the TS smoothens or sharpens apparent variability in density classified by a natural breaks algorithm (Fig. 6). In this particular case, this corroborates the observation made above, but more generally, it is evident that this can give up differential spatial patterning in find distribution maps.

Translating back to total find densities in the TS, in 2320 there is a numeric density of 1.3 sherds/m^2 with a weight density of 37 g/m^2, where in 2322 there is a numeric density of 0.9 sherds/m^2 with a weight density of 16 g/m^2. Therefore, where calculated in counts, there is a $1.3/0.9 = 1.4$ factor difference (FD) (Pettegrew, 2014) in density, when calculated in weights there is a $37/16 = 2.3$ FD in density. The degree to which relative densities scale up or down for counts and weights is quite different. As a final example, for our comparison between sample types, the PS gave up almost three times as many PW sherds when counting, but only two times as many PW sherds when weighting, evidently the result of picking up smaller finds. Clearly such differences can affect interpretations, as the information of the surface distributions noticeably changes looking at either numbers or weights.

CONCLUSIONS

This elaborate assessment has demonstrated the importance of a careful treatment of breakage and the possible biases as a result of it for archaeological field survey. With a weight variation that is partly the result of breakage as variable as suggested in the case study, weight should be taken into account as an estimator for abundance. Weight has its specific problems, and there is the issue of variation caused by vessel size range. To get a good grip on quantitative patterns, it is imperative to assess counts and breakage next to weight, and study spatial patterning in them. The integrative assessment of these basic quantifiers for surface assemblages is imperative because it may also provide information about states of conservation and formation in various places. Furthermore, a more intensive sampling technique results in more and smaller pieces of pottery is a warning for assessing numeric abundance in research designs where various sampling methods are combined.

As said, with weight there are problems such as the aforementioned vessel size range and specific weight, which probably generate noise, or worse, introduce notable error in weight densities. Although such variability can be great due to the broad chronological frame of the CW and PW wares, that potential bias is mitigated due to the actual finds under study largely dating to the fourth–third centuries BCE. Nevertheless, it is sensible to assess the variability they can exhibit and consider those as error margins on the weight densities. A way forward here, and something really needed, is to design empirical studies targeting vessel size ranges and specific weight for wares to better understand potential influence. However, it is important for examinations such as these to be able to assess the finds at the level of the individual object. Since this takes a lot of effort, a statistical subsampling design should be applied. Alternatives for weight such as volume are practically cumbersome, though possibly the fast development of three-dimensional recording can tip the balance in favour of such an approach in the near future.

ACKNOWLEDGEMENTS

This study was executed in the context of the Tappino Area Archaeological Project (TAAP), coordinated by Dr Tesse D. Stek, University of Groningen and the Koninklijk Nederlands Instituut in Rome. The TAAP is organized by Tesse Stek, Rogier A. A. Kalkers, Dr Jesús García Sánchez and the present author. The work at Colle S. Martino was conducted with permission of the Soprintendenza Archeologia, Belle Arti e Paesaggio del Molise. The fieldwork would have been impossible without the support provided by the Comune di Jelsi, especially Michele Fratino, and our collaborative research centre in Molise: the Centro Didattico di Studi Archeologici di Jelsi (CeDISA), and the team itself: at Colle S. Martino consisting of Tesse Stek, Rogier Kalkers, Lennart Kruijer, Koos Mol and Max Caspers in the field, and Lisa Götz, Lucia Lecce, Sheila Cherubini, Filippo Salamone, Jacqueline Splinter and Rogier Kalkers for finds analyses. Particular thanks to readers of the first drafts of this study, that is, Professor Dr Vladimir V. Stissi, Tesse Stek and Professor Dr James Symonds, Jesús García Sánchez and Rogier Kalkers.

PEER REVIEW

The peer review history for this article is available at https://publons.com/publon/10.1111/arcm.12720.

REFERENCES

Adroher Auroux, A. M., Carreras Monfort, C., De Almeida, R., Fernández Fernández, A., Molina Vidal, J., & Viegas, C. (2016). Record for the quantification of archaeological pottery: State of art and a new proposal. Seville protocol (PRCS/14). *Zephyrus*, *78*, 87–110. https://doi.org/10.14201/zephyrus20167887110

Akkeraz, A., & Collins-Elliot, S. A. (2017). Gardens of the Hesperides: The Rural Archaeology of the Loukkos Valley. In *Interim report on the 2016 season* (pp. 2017–2015). FOLDER Survey.

Alcock, S., Cherry, J., & Davis, J. (1994). Intensive survey, agricultural practice and the classical landscape of Greece. In I. Morris (Ed.), *Classical Greece: Ancient histories and modern archaeologies* (pp. 137–170). Cambridge University Press.

Ammerman, A. (1985). Plow-zone experiments in Calabria, Italy. *Journal of Field Archaeology*, *12–1*, 33–40. https://doi.org/10.2307/529373

Arcelin, P., & Tuffreau-Libre, M. (Eds.) (1998). *La quantification des céramiques: conditions et protocole: actes de la table ronde du Centre archéologique européen du Mont-Beuvray (Glux-en-Glenne, 1998)*. Centre archéologique européen du Mont Beuvray, Glux-en-Glenne.

Banning, E. B. (2002). *Archaeological survey*. Kluwer Academic Publishers.

Barker, G. (Ed.) (1995). *The Biferno Valley survey: The archaeological and geomorphological record*. Leicester University Press.

Barton, C. M., Bernabeu, J., Emili Aura, J., & Garcia, O. (1999). Land-use dynamics and socioeconomic change: an example from the Polop Alto Valley. *American Antiquity*, *64–4*, 609–634. https://doi.org/10.2307/2694208

Baumhoff, M. A., & Heizer, R. F. (1959). Some unexploited possibilities in ceramic analysis. *Southwestern Journal of Anthropology*, *15*, 308–316. https://doi.org/10.1086/soutjanth.15.3.3628980

Bevan, A., & Conolly, J. (2004). GIS, archaeological survey, and landscape archaeology on the island of Kythera, Greece. *Journal of Field Archaeology*, *29–1*(2), 123–138. https://doi.org/10.1179/jfa.2004.29.1-2.123

Bintliff, J., Farinetti, E., Slapšak, B., & Snodgrass, A. (2017). Boeotia Project. In *The City of Thespiai. Survey at a complex urban site* (Vol. 2). McDonald Institute for Archaeological Research.

Bintliff, J., Howard, P., & Snodgrass, A. (1999). The hidden landscape of prehistoric Greece. *Journal of Mediterranean Archaeology*, *12–2*, 139–168. https://doi.org/10.1558/jmea.v12i2.139

Blanton, R. E. (2001). Mediterranean myopia. *Antiquity*, *75*, 627–629. https://doi.org/10.1017/S0003598X00088918

Byrd, J. E., & Owens, D. D. (1997). A method for measuring relative abundance of fragmented archaeological ceramics. *Journal of Field Archaeology*, *24*, 315–320. https://doi.org/10.1017/S0003598X00088918

Carrete, J. M., Kaey, S. J., & Millett, M. (1995). *A Roman Provincial Capital and its Hinterland, the survey of the territory of Tarragona, Spain, 1985–1990* (Vol. 15). JRA supplementary series. University of Michigan.

Chadwick, A. M., & Evans, H. (2000). Reading Roystone's rocks: Landscape survey and lithic analysis from test pitting at Roystone grange, Ballidon, Derbyshire, and its implications for previous interpretations of the region. *Derbyshire Archaeological Journal*, *120*, 101–122. https://doi.org/10.5284/1038992

Coccia, S., & Mattingly, D. (1992). Settlement history, environment and human exploitation of an intermontane basin in the central Apennines: The Rieti survey 1988–1991, part I. *Papers of the British School at Rome*, *60*, 213–289. https://doi.org/10.1017/S0068246200009831

Dawson, G. J. (1971). *Montague Close Part 2* (Vol. 1, pp. 250–251). London Archaeology.

Dunnell, R. C., & Dancey, W. S. (1983). The Siteless survey: A regional scale data collection strategy. *Advances in Archaeological Method and Theory*, *6*, 267–287. https://doi.org/10.1016/B978-0-12-003106-1.50012-2

Dunnell, R. C., & Simek, J. F. (1995). Artifact size and Plowzone processes. *Journal of Field Archaeology*, *22*, 305–319. https://doi.org/10.2307/530178

Egloff, B. J. (1973). A method for counting ceramic rim sherds. *American Antiquity*, *38*, 351–353. https://doi.org/10.2307/279724

Fentress, E. (2000). What are we counting for? In R. Francovich & H. Patterson (Eds.), *Extracting meaning from Ploughsoil assemblages* (pp. 44–52). Oxbow Books.

Foley, R. (1981). Off-site archaeology: an alternative approach for the short-sited. In *Pattern of the past: Studies in honour of David Clarke* (pp. 157–183). Cambridge University Press.

Fulford, M. G. (1973). The excavation of three Romano-British pottery kilns in Amberwood Enclosure, near Fritham, New Forest. *Proceedings of the Hampshire Field Club and Archaeological Society*, *28*, 5–27.

Gallant, T. W. (1986). "Background noise" and site definition: A contribution to survey methodology. *Journal of Field Archaeology*, *13–4*, 403–418. https://doi.org/10.1179/jfa.1986.13.4.403

Given, M. (2004). Mapping and manuring: Can we compare sherd density figures. In *Side-by-side survey: Comparative regional studies in the Mediterranean world* (pp. 13–21). Oxbow Books.

Hinton, P. A. (1977). Rudely made earthen vessels of the twelfth to fifteenth centuries A.D. In D. P. S. Peacock (Ed.), *Pottery and early commerce: Characterization and trade in Roman and later ceramics* (pp. 221–238). Academic Press.

Hulthén, B. (1974). On choice of element for determination of quantity of pottery. *Norwegian Archaeological Review*, *7*, 1–5. https://doi.org/10.1080/00293652.1974.9965196

Kinnunen, J. (2020). Weight or density corrected value? Using density derived key ratio for additional accuracy to Intercomparability of medieval and historical artifact groups. *International Journal of Historical Archaeology*, *24*, 62–78. https://doi.org/10.1007/s10761-019-00503-0

Kintigh, K. W. (1988). The effectiveness of subsurface testing: A simulation approach. *American Antiquity*, *53*, 686–707. https://doi.org/10.2307/281113

Krakker, J. J., Shott, M. J., & Welch, P. D. (1983). Design and evaluation of shovel-test sampling in regional archaeological survey. *Journal of Field Archaeology*, *10*(4), 469–480. https://doi.org/10.1179/009346983791504147

Krijnen, A. L., Waagen, J., & Hilditch, J. R. (Accepted/In press). Survey, ceramics and statistics: the potential for technological traits as chronological markers. In A. Meens, M. Nazou, & W. van de Put (Eds.), *Fields, Sherds and scholars: Recording and interpreting survey ceramics*. Sidestone Press.

Lightfoot, K. G. (1989). A defense of shovel-test sampling: A reply to Shott. *American Antiquity*, *54*, 413–416. https://doi.org/10.2307/281716

Mateo Corredor, D., & Molina Vidal, J. (2016). Archaeological quantification of pottery: The rims count adjusted using the modulus of rupture (MR). *Archaeometry*, *58*, 333–346. https://doi.org/10.1111/arcm.12171

Millet, M. (1991). Pottery: population or supply patterns? The Ager Tarraconensis approach. In *Roman landscapes: Archaeological survey in the Mediterranean region* (pp. 18–29). British School at Rome.

Millett, M. (1979). How Much pottery. In M. Millett (Ed.), *Pottery and the archaeologist* (pp. 77–80). Routledge.

Millett, M. (2000). Dating, quantifying and utilizing pottery assemblages from surface survey. In R. Francovich & H. Patterson (Eds.), *Extracting meaning from Ploughsoil assemblages* (pp. 53–59). Oxbow Books.

Molina Vidal, J. (1997). *La dinámica comercial romana entre Italia e Hispania Citerior (siglos II a. C. – II d. C.)*. Universidad de Alicante, Instituto de Cultura Juan Gil-Albert, Alicante.

Odell, G. H., & Cowan, F. (1987). Estimating tillage effects on artefact distributions. In *American antiquity 52–3* (pp. 456–484). Cambridge University Press.

Orton, C. (1975). Quantitative pottery studies: Some progress, problems and prospects. *Science and Archaeology*, *16*, 30–35.

Orton, C. (1982). *Mathematics in archaeology*. Cambridge University Press.

Orton, C. (1993). How Many Pots Make Five—An Historical Review Of Pottery Quantification. In *Archaeometry 35-2* (pp. 169–184). Wiley-Blackwell.

Orton, C. (2000). *Sampling in archaeology*. Cambridge University Press.

Orton, C., & Tyers, P. (1993). Counting broken objects: The statistics of ceramic assemblages. *Proceedings of the British Academy*, *77*, 163–184.

Orton, C., Tyers, P., & Vince, A. (2013). *Pottery in Archaeology* (2nd ed.). Cambridge University Press.

Palumbo, S. (2015). Assessing the utility of plowed field surface deposits in pilot research. *Advances in Archaeological Practice*, *3*(1), 78–92. https://doi.org/10.7183/2326-3768.3.1.78

Pelgrom, J., & Stek, T. D. (2010). A landscape archaeological perspective on the functioning of a rural cult place in Samnium: Field surveys around the sanctuary of S. Giovanni in Galdo (Molise). *Journal of Antique Topography*, *XX, 41–102*.

Pettegrew, D. K. (2014). Chapter 3—Survey Data and Experiments in Sampling. In W. Caraher, R. S. Moore, & D. K. Pettegrew (Eds.), *Pyla-Koutsopetria I archaeological survey of an ancient coastal town*. The American Schools of Oriental Research.

Poulain, M. (2013). Notes on the quantification of post-medieval pottery in the Low Countries. *Post-Medieval Archaeology*, *47*(1), 106–118. https://doi.org/10.1179/0079423613Z.00000000027

Py, M. (1991). Système d'enregistrement, de gestion et d'exploitation de la documentation issue des fouilles de Lattes. In *Lattara 4*. Lattes.

Raux, S. (1998). Méthodes de quantification du mobilier céramique. Etat de la question et pistes de réflexion. In P. Arcelin & M. Tuffreau-Libre (Eds.), *La quantification des céramiques: conditions et protocole: actes de la table ronde du Centre archéologique européen du Mont-Beuvray (Glux-en-Glenne, 1998)* (pp. 11–16). Centre archéologique européen du Mont Beuvray.

Rice, P. M. (1987). *Pottery analysis: A source book*. University of Chicago Press.

Slane, K. W. (2003). Corinth's Roman pottery: Quantification and meaning. *Corinth*, *20*, 321–335. https://doi.org/10.2307/4390731

Stek, T. D. (2018). Exploring non-urban society in the Mediterranean: Hill-forts, villages and sanctuary sites in ancient Samnium, Italy. *Antiquity*, *92*, 1–7. https://doi.org/10.15184/aqy.2018.155

Stek, T. D., & Pelgrom, J. (2005). Samnite sanctuaries surveyed: Preliminary report of the sacred landscape project 2004. *BABesch*, *80*, 65–71. https://doi.org/10.2143/BAB.80.0.630018

Stek, T. D., & Waagen, J. (forthcoming). Scratching the surface. Integrating low-visibility zones and large rural sites in landscape archaeology using point sampling: a quantitative comparison with standard Mediterranean field survey methods in on- and off-site conditions.

Strack, S. (2011). 'Erfahrungsbericht' of application of different quantitative methods at Kalapodi, Sara Strack. In S. Verdan, T. Theurillat, & A. K. Pfyffer (Eds.), *Early iron age pottery: A quantitative approach (BAR-IS 2254)* (pp. 23–38). Oxford.

Taylor, J. (2000). Cultural depositional processes and post-depositional problems. In R. Francovich & H. Patterson (Eds.), *Extracting meaning from Ploughsoil assemblages* (pp. 16–26). Oxbow Books.

Terrenato, N. (2004). Sample Size Matters! The paradox of global trends and local surveys. In S. Alcock & J. F. Cherry (Eds.), *Side-by-side survey* (pp. 36–48). Oxbow Books.

Tol, G. W. (2012). *A fragmented history. A methodological and artefactual approach to the study of ancient settlement in the territories of Satricum and Antium*. Barkhuis and Groningen University Library.

van Leusen, P. M. (2002). *Pattern to process, methodological investigations into the formation and interpretation of large-scale patterns in archaeological landscapes. Ph.D. dissertation*. University of Groningen.

Verdan, S. (2011). Pottery quantification: Some guidelines. In *In early iron age pottery: A quantitative approach, proceedings of the international round table organized by the Swiss School of Archaeology in Greece (Athens, November 28–30, 2008)* (pp. 165–171). Archaeopress.

Waagen, J. (2014). Evaluating background noise: Assessing off-site data from field surveys around the italic sanctuary of S. Giovanni in Galdo, località Colle Rimontato, Molise, Italy. *Journal of Field Archaeology*, *39–4*, 417–429. https://doi.org/10.1179/0093469014z.00000000099

Winther-Jacobsen, K. (2010). From pots to people. In *A ceramic approach to the archaeological interpretation of ploughsoil assemblages in late Roman Cyprus* (Vol. 17). BABesch Supplement.

Witcher, R. (2006). Broken pots and meaningless dots? Surveying the rural landscapes of Roman Italy. *Papers of the British School at Rome*, *74*, 39–72. https://doi.org/10.1017/S0068246200003226

How to cite this article: Waagen, J. (2022). Breakage, bias and the archaeological surface record: Assessing the quantification problem in archaeological field survey. *Archaeometry*, *64*(2), 529–544. https://doi.org/10.1111/arcm.12720