Journal of Archaeological Science 128 (2021) 105346

Contents lists available at ScienceDirect

Journal of Archaeological Science

journal homepage: http://www.elsevier.com/locate/jas

# Detecting variability: A study on the application of bayesian multilevel modelling to archaeological data. Evidence from the Neolithic Adriatic and the Bronze Age Aegean

Christianne L. Fernée *, Konstantinos P. Trimmis

*Department of Anthropology and Archaeology, University of Bristol, UK*

ARTICLE INFO

*Keywords:*
Variability
Archaeological data
Multilevel modelling
Bayesian statistics in archaeology
Akrotiri Thera
Neolithic Adriatic

ABSTRACT

The detection and interpretation of variability in archaeological data has been a long-standing effort in the field. This paper aims to introduce the application of Bayesian multilevel modelling as a tool for the detection of variability at levels within nested archaeological data. Model structure, ways of construction, and the potential of using variability information to enhance archaeological interpretations is presented. This is demonstrated through the analysis of two case study datasets: Neolithic pottery finds from Mala (Nova) Pećina cave excavations in Croatia and stone finds from the Bronze Age site of Akrotiri, Thera, Greece. This is followed by a discussion of the multilevel model results and the possible interpretations that can be derived from them. Finally, propositions are made on how these and other models can be extended.

## 1. Introduction

The idea of using statistics and computational modelling for the analysis of archaeological datasets appeared with the emergence of New Archaeology in the late 1960s (Doran and Hodson 1975; Drennan 2009; Orton 1980; Sullivan and Olszewski 2016; White and Thomas 1972 and more). Since then researchers have used an array of methods to detect and interpret variability in archaeological datasets. To name a few, methods include the use of Coefficient of Variation (CV) on lithic artefact patterns (see a review on Garvey 2018), bivariate regression analysis on ethnoarchaeological data to understand human mobility (Kent 1992), Random-effects Logistic Regression Analysis to study agricultural practices (McCorriston 2002), and hierarchical classification systems and cluster analysis on archaeological ceramics (Plog 1980) (see more on variability in archaeological data in O'Shea 1984; Roberts and Van der Linden 2011; Sullivan and Olszewski 2016; Schiffer and Skibo 1997).

Gradually, from the late 1970s through the 1980s, the endeavour of New Archaeology to explain the past through purely quantitative, computerised, statistical, and analytical methods received scepticism by several authors (e.g. Earle and Preucel 1987; Hodder 1986; Hole 1980; Hurst Thomas 1978; Shanks and Tilley 1982). The consequent rise of interpretational approaches to archaeological data brought more methods of detecting variability and understanding its meaning in archaeological assemblages. These methods were distant from previous statistical and computational analyses of datasets; they were experimental, typological and cognitive (see Sullivan and Olszewski 2016). However, as suggested by Hurst Thomas, 'trends in research swing back and forth like a pendulum' (1978: 240), thus meticulous analysis of archaeological datasets through applications of statistical and computational modelling are again the norm. In the last decade the interpretational 'gap' between the descriptive and testing statistical applications has been reconciled. This has, in part, been a consequence of the increasing application of Bayesian Statistics into archaeological practice, which incorporates the researcher's beliefs, and to an extent the researcher's perspective, into the statistical analysis (for a review see Otarola-Castillo and Torquato 2018).

Existing methods used to study variation within archaeological data, including those mentioned above, analyse variation at a single level at a time. This can be either geographical or temporal, such as regional or diachronic variation using CV and regressions (Garvey 2018; Liu et al., 2020; Schmid 2019) or variation over time using time-series analysis (Gayo et al., 2015). However, archaeological data is inherently multi-layered and these methods fail to exploit this fully. Previously used methods are limited by the inability to concurrently analyse variation in archaeological data at its different levels, such as site within a

* Corresponding author.
*E-mail address:* christianne.fernee@bristol.ac.uk (C.L. Fernée).

https://doi.org/10.1016/j.jas.2021.105346
Received 4 February 2020; Received in revised form 11 February 2021; Accepted 11 February 2021
Available online 6 March 2021

region, without violating the assumptions that underly them. A method that overcomes this limitation is multilevel modelling. Recently, Bayesian mutlievel models have been applied to archaeological data. They have been used for regional chronological sequences (Banks et al., 2019), to analyse variability in isotopic signatures (Perri et al., 2019) and to examine element- and assemblage-level variation in biometric measurements (Wolfhagen 2020).

## 2. Aims, materials and methods

This paper introduces multilevel modelling as a tool for the detection and interpretation of variability archaeological datasets, going beyond previous applications of Bayesian modelling in archaeology. Despite the nested nature of archaeological data (e.g. an artefact or individual, within a site, within a region) multilevel modelling it is yet to be fully exploited within archaeology. Multilevel models (MLMs) can be used to answer a variety of questions, regarding both the levels in the respective model and associated predictors. Problematic questions can be answered by increasing the models' complexity. It is also easily accessible and can be carried out using specific computer software such as MLwiN (Charlton et al., 2020; Browne 2019) or using packages available in open source code statistical software environments such as R (R Core Team 2018). These packages include R2MLwiN (Zang et al., 2016), which uses MLwiN, lme4 (Bates et al., 2015), which used traditional maximum likelihood estimation, and rjags (Plummer et al., 2019), rstan (Stan Development Team 2020b), brms (Bürkner 2017) and cmdstanR (Stan Development Team 2020a) that use Bayesian estimation methods.

Multilevel modelling can be applied to an array of archaeological data to understand where variations occur (see Ferneé 2020). It can be applied to different types of excavation data – pottery, lithics, bones etc – to detect material variations between artefacts, contexts, trenches, sites, or regions, dependent on the number of levels constructed. Once variation has been partitioned archaeological interpretations can be drawn. Interpretations can vary in complexity depending on the model structure.

The aim of this paper is to present the nature of MLMs and to introduce their use for the analysis of archaeological assemblages. In order to achieve this aim, a feasibility study on the usefulness of multilevel modelling in archaeological interpretation has been designed and showcased by the two separate datasets. The first dataset involves a Neolithic pottery assemblage from Eastern Adriatic and the second a Bronze Age worked stone assemblage from the Aegean.

### 2.1. Multilevel modelling

Multilevel models, also known as random coefficient models and hierarchical models, are used on data that is nested/hierarchical in nature. They are used ever increasingly in social, biological and medical sciences (Kim et al., 2018; O'Malley et al., 2014; Brunton-Smith and Sturgis 2011). They have been used for a variety of goals including causal inference, prediction and descriptive modelling (Gelman and Hill 2007). In archaeology, MLMs can be used wherever hierarchical datasets appear, for example an artefact category, such as pottery sherds or lithics, in a context, in a trench, in an excavation site, in a region. They can overcome the limitations of methods currently used in archaeology, such as CV and regressions, that are unable to study variation at multiple levels concurrently.

Multilevel models are extensions of regressions, in which data are structured into groups and coefficients can vary by group. However, rather than using a single large data matrix, a matrix is constructed for each level within the hierarchy (Gelman and Hill 2007). The recognition of a hierarchy in the data allows for the assessment of variation at each level. For example, a two-level model of pottery sherds within excavation contexts would allow the assessment of variation between contexts and between pottery sherds within contexts. This enables the researcher to understand where and how effects are occurring (Kharazifard et al., 2017).

Performing an analysis that does not recognise the presence of nested data, such as those that are traditionally used in archaeology, will create problems with intraclass correlations, chi-squared statistics, parameters and the underestimation of their standard errors. Type 1 errors become more frequent; predictors appearing to have a significant effect when they do not (Steenbergen and Jones 2002). The specific group configurations alongside between-group and within-group structures compound problems (Julian 2001). The duplication of observations violates the independence assumptions of traditional analyses such as ANOVA and OLS regression (Steenbergen and Jones 2002; Gelman and Hill 2007). This duplication of observations is not an issue in MLMs.

A common exercise in archaeology is the analysis of differences in an artefact feature/variable across multiple archaeological sites. For example, analysing differences in pottery thickness or waretype across different types of archaeological sites (e.g. flat settlement, cave, tell, in a Near Eastern context). Differences are often analysed without acknowledging the nested nature of this data using, for example, an ANOVA for continuous data or a chi-squared test for categorical data. Failure to consider the levels in data, such as pottery sherds within a context within a trench within an archaeological site, mean that significant differences that are identified may not actually exist, which can in turn result in incorrect inferences and conclusions being drawn.

MLMs are, however, based on three main assumptions: linear relationships, homoscedasticity and normal distribution of residuals (Maas and Hox 2004). However, heteroscedasticity can be modelled directly to account for violations of homoscedasticity (Goldstein 1995:48–57). Likewise, multilevel estimation methods have been found to be robust against violations of normality of residuals at the second level (Maas and Hox, 2004). The limitations of MLMs are largely the same as other statistical methods, for example, if samples are too small they do not provide an adequate basis for statistical inferences and their results, such as estimations of predictor effects, do not necessarily show causality (Gelman 2006). They help describe, summarise and quantify patterns in the data, however they do not explain these patterns, they require careful interpretation by the researcher. Finally, it is an advanced statistical technique that requires are good grounding in statistics, however resources are becoming increasingly available.

#### 2.1.1. Model structure

Multilevel modelling explores hierarchical data structures. In these structures' units are grouped at different levels, level 1 is the lowest level which is nested within level 2. For example, pottery sherds may be the level 1 unit nested within excavation, which are the level 2 unit (Fig. 1). This structure can also be used for repeated measures in longitudinal studies. For example, repeated measures over time would be the level 1 unit nested within the specimen measured, which are the level 2 unit.

Levels are 'random' as they have been sampled from a wider population (although when this is not possible a more general motivation is suitable, see Gelman et al., 2013). For example, in the model described above, pottery sherds would have been sampled from a wider-number of sherds within a context, and the contexts would have been sampled from a wider population in a site (Fig. 2). Like traditional regression analysis, explanatory units, known as predictors, can be input into an MLM. Predictors and levels are distinguished by their nature. Predictors are 'fixed' variables or effects that have a small number of fixed categories. These fixed effects are parameters that do not vary, and their observations are independent. For example, in the pottery model described above the type of decoration (decoration or no decoration) can be input as a predictor at level 1 and the context type (pit, trench, posthole) can be input as a predictor at level 2 (Fig. 2).

Random effects are grouping variables that are clustered. Grouping variables, unlike fixed effects, are non-independent. Therefore, observations that are non-independent should be specified as random effects, such as pottery sherds within contexts. In MLMs, random effects estimate the variance between groups rather than the mean of each group.

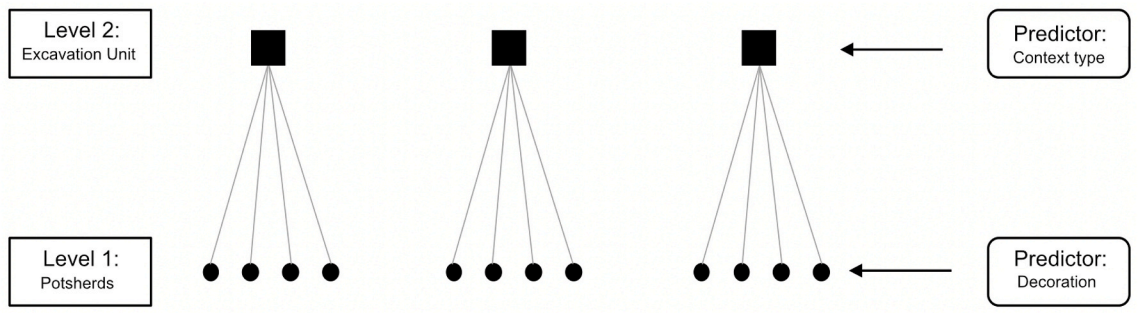


**Fig. 1.** Diagram of an example multilevel model structure. Pottery sherds (circles) are the level 1 units grouped into excavation units (squares) at level 2. The model also includes fixed explanatory predictor variables: sherd decoration at level 1 and context type at level 2.

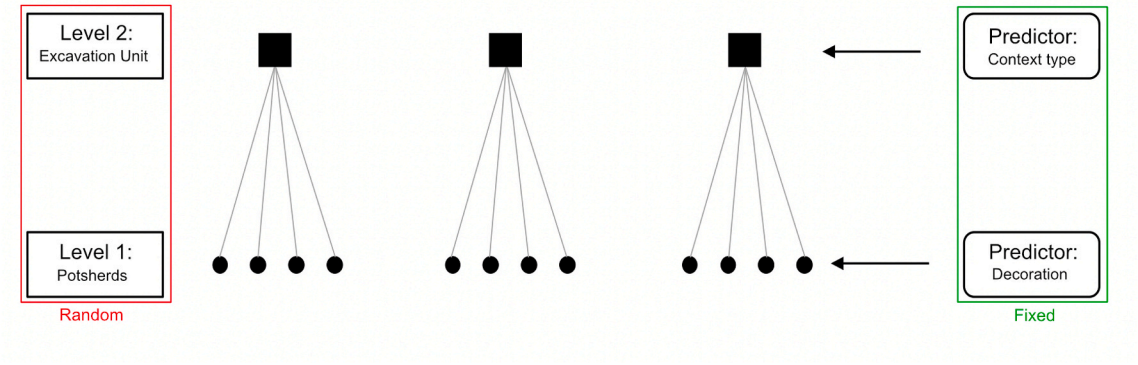


**Fig. 2.** Example of a multilevel model structure highlighting the random (red) and fixed (green) components. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

Statistical comparisons of means more commonly come from fitting variables as fixed effects (Theobald, 2018).

Random effects can be specified in a multilevel model in three ways depending on the relationship between variables: 1) a random intercept, 2) random slopes and 3) random slopes and intercept (Theobald, 2018). A random intercept model allows each grouping variable to have its own intercept but equivalent slopes; this can be interpreted as the average of each group is different but the relationship with the predictors is the same (Fig. 3a). In the pottery model example, this would allow the outcome variable, pottery thickness, of each group, excavation unit, to have its own mean whilst all units have the same relationship with predictor, decoration. A random slopes model allows the relationship between the predictors and the outcome variable to vary but the starting point, or the outcome mean, is the same (Fig. 3b). For the pottery model example, the mean pottery thickness of each excavation unit will be held at the same point whilst the relationship between pottery thickness and decoration will be allowed to vary. Finally, random slopes and intercept models each group is allowed a unique intercept and a unique relationship with predictors (Fig. 3c). In the model example, both the mean pottery thickness of each excavation unit and it's relationship with sherd decoration would be allowed to vary.

### 2.1.2. *Estimation methods in multilevel models*

The contribution of the 'random' and 'fixed' components of the model are commonly estimated using Maximum Likelihood (ML), Restricted Maximum Likelihood (REML) or Iterative Generalized Least Squares (IGLS) algorithms. ML and REML are commonly used, but issues arise when using them (see El-Hobarty et al., 2018) and IGLS can be biased with for smaller group sizes (Goldstein 2002). Bayesian Markov Chain Monte Carlo (MCMC) estimation procedures provide a more robust estimation than traditional IGLS methods (Browne and Jones, 2006; Jones & Subrmanian 2013, 2017).

The Bayesian fitting of an MLM requires, as usual in Bayesian statistics, a prior distribution for the parameters. The most commonly used

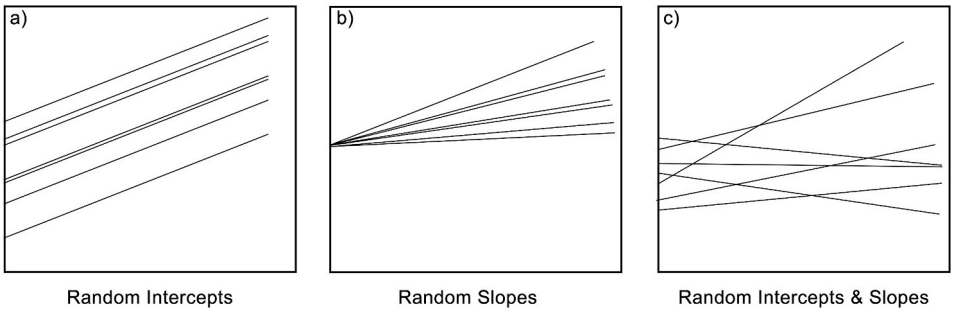


**Fig. 3.** Types of random effects in multilevel model: a) a random intercepts model, where intercepts the slopes are the same for each group but the intercepts vary, b) A random slope model, where the slopes vary for each group but the intercepts are the same and c) a random intercepts and random slopes model, where both the slopes and intercepts vary by group.

priors when fitting MLMs are either informative or uninformative. Informative priors are distributions express a strong belief about a parameter and will strongly influence the posterior distribution and conclusions. Consequently, many statisticians prefer uninformative or diffuse priors which have little influence on the conclusions (Hox 2010). More detailed and extensive overviews of Bayesian methods for the analysis of MLMs can be found elsewhere (Gelman et al., 2007; Gelman et al., 2013 Gelman et al., 2017; Browne 2019).

MCMC methods are simulation-based procedures that make a high number of iterations or a chain. Chains are initiated at a particular starting value and can take a while to settle on a posterior distribution, known as converging. This period of time is known as the burn-in period which is omitted from the sample summaries (Browne 2019, 5). This is followed by the monitoring period which are the iterations that are saved and used for posterior inference.

In Bayesian MLMs, a Deviance Information Criterion (DIC) can be used to identify the best fitting model whilst penalising increasing complexity (Spiegelhalter et al., 2002). A decrease in DIC indicates a 'better' model (Browne 2019, 234). Models with a difference of 2 or less should be considered along with the best model, 4–7 difference have considerably less application and 10 or more the model can be omitted from consideration (Jones and Subrmanian 2017). The DIC should also be considered alongside changes in estimates of variance and the posterior mean and standard deviation of the regression (fixed) coefficients (Dias et al., 2011).

The MCMC posterior means and 95% credible intervals can be presented as summaries of the posterior distribution for each variable. The 95% credible interval is a Bayesian analogue of traditional 95% confidence intervals. Yuan and Mackinnon (2009) suggest that Bayesian credible intervals are more meaningful and relevant to scientific practice. This is because credible intervals have natural interpretation: the true value is contained within a 95% credible interval with 95% (posterior) probability. A 95% credible interval contains 95% of the posterior density.

#### 2.1.3. Variable analysis

A Variance Partition Coefficient (VPC) can be calculated to determine the proportion of random variance. When predictors are included, a model can be separated into fixed and random components; the levels are the random component and the predictors are the fixed component (Fig. 2). Residual variance occurs in the random part of the model, which can be proportioned for each level. For a two-level model this would be the proportion of residual variation occurring at level 2. The VPC reflects the proportion of the residual variance that is due to differences between groups (Goldstein 2010). According to the pottery example, the VPC would reflect the variance that is due to differences between contexts. The manner by which this is calculated is dependent on the nature of the variable under analysis. The random effects can be assessed visually through plotting the residuals. Residuals reflect the departures of groups at each level from the overall mean, the intercept, which allow comparisons to be made between groups. Therefore, residuals can be used to inspect group effects at each level (Fig. 4).

Fixed variables can be analysed in various ways. Significance testing can be used to test the difference within each fixed parameter. Sensitivity analysis can also be used: this is the study of how variation in the output of the model can be apportioned to the different sources of variation (Saltelli 2008, 3). In an MLM, it can be used to understand how explanatory variables contribute to the model, particularly determining which parameter contributes most to the variability of the dependent variables.

### 2.2. Continuous and categorical response models

Continuous data can easily be fitted to traditional MLMs. For an MLM fitted to continuous data, the VPC is equal to the intra-unit correlation, which is the correlation between level 1 units in the same level 2 unit (Rasbash et al., 2017). For example, sherds within the same context. They can also be calculated in three- and higher-level models as well as in models with more complex random effect structures (e.g. cross-classified, multiple membership, spatial, and dyadic structures) (Leckie et al., 2019). They are calculated by dividing the level 2 residual variance by the combined level 2 and level 1 residual variance (Fig. 5).

Unlike models for continuous response variables, multilevel models for binary or proportion responses largely use a logistic link function. It has been shown MCMC methods with diffuse priors are less biased than quasi-likelihood methods for binary response models (Browne et al., 2005). When fitting models for categorical (binary, ordinal, or nominal) and count responses partitioning variance is more challenging to calculate than in models with a continuous response variable, where it is reasonably straightforward to calculate (Leckie et al., 2019). This is due

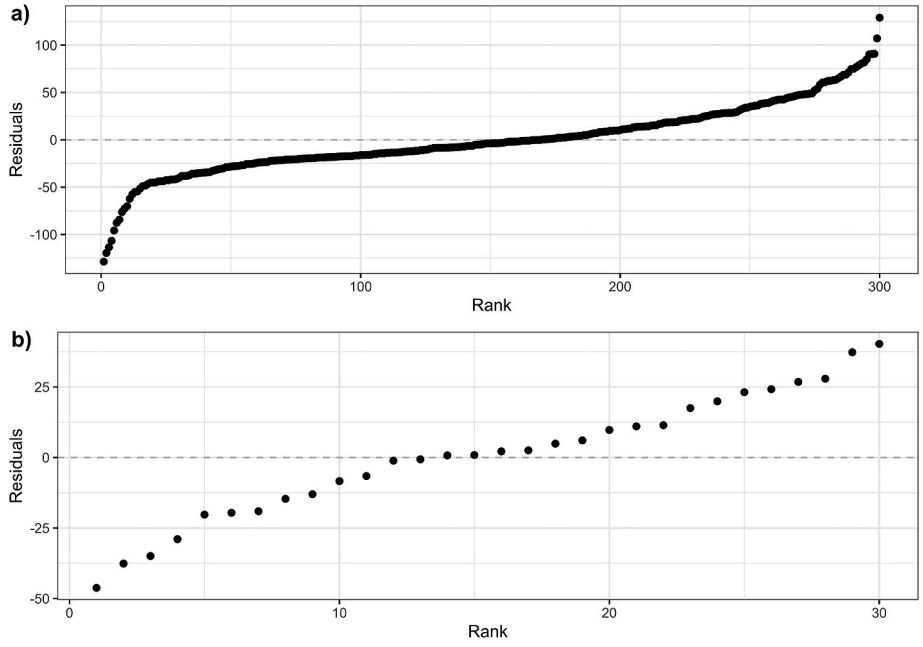


**Fig. 4.** Example residual plots for a 2-level multilevel model: a) Level 1 and b) Level 2.

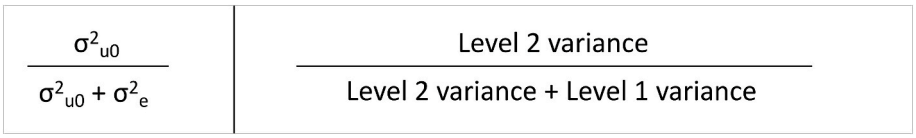


**Fig. 5.** Calculation of VPC for continuous response models: formula (left) and general explanation (right).

to a number of issues, this includes that individual and cluster components of variance are calculated on difference scales, the discrete response scale and linear predictor scale respectively. In addition to this the components of variance depend on the covariates, so there is typically no unique VPC for models with discrete outcomes (Austin et al., 2017). Goldstein (2002) considered four approaches for estimating the VPC in a binary response models, three of which were later used by Browne et al. (2005). The four approaches are: model linearization, simulation, a binary linear model and a latent variable approach (Goldstein et al., 2002). Leckie et al. (2019) found that although VPC estimates differ slightly when using the different methods the interpretation of them remains the same, overall they are indicative of whether contextual effects a large or small.

The most commonly used method to report VPCs in categorical responses is the latent variable approach. This approach views the observed categorical variable as arising from an underlying continuous variable e.g. a continuous exam score scale underlying the observed binary pass or fail status (Leckie et al., 2019; Goldstein et al., 2002). The true underlying variable is continuous but we can only observe a binary response that indicates whether the underlying variable is greater or less than a given threshold (Browne et al., 2005). This approach assumes that the level 1 variance is fixed and independent of the predictor variables (Browne et al., 2005). This often appealing as it allows VPCs to be calculated for categorical responses using essentially the same expressions as those derived for continuous responses (Leckie et al., 2019). However, it is assumed that in the logistic regression model the underlying variables will come from a logistic distribution, with a variance of $\pi^2/3$ ($\pi^2/3 \approx 3.29$) (Browne et al., 2005). This is substituted for the level 1 variable, resulting in the formula in Fig. 6.

However, unlike the other methods, the simulation-based method does not just give an approximation (Browne et al., 2005). Simulation-based methods do not rely on the same assumptions as the latent variable approach, rather they are dependent on specific covariate patterns (for details of method see Browne et al., 2005; Goldstein et al., 2002; Rasbash et al., 2019). Therefore, it is possible that different VPC values could be obtained for each distinct covariate pattern (Leckie et al., 2019). The advantages of the simulation method it is more accurate as it does not rely on approximations and is simple and fast to compute (Goldstein 2002; Browne et al., 2005). However, it can become more time consuming and difficult to calculate when more complex models are studied with more than two levels (Browne et al., 2005).

## 3. Applying multilevel modelling to archaeological data

To showcase the feasibility of the application of MLMs to archaeological data, variation is studied in two very different archaeological datasets. First, the stone 'sphere' assemblage found during excavations at the Bronze Age town of Akrotiri on Santorini, Greece (see Tzachili 2007 and Valacy forthcoming). This is followed by the Neolithic pottery assemblage from the Mala (Nova) Pećina cave in Croatia, unearthed during the 2016 excavation season (see Drnić et al., 2018; Trimmis and Drnić 2018).

For each dataset MLMs were constructed and analysed in MLwiN 3.05 (Charlton et al., 2020; Browne 2019). A Bayesian MCMC estimation procedure, using Gibbs sampling, was used with the default MLwiN diffuse inverse-gamma (0.001, 0.001) prior distribution for the random effects and an inproper uniform prior of $\propto 1$ for the fixed effects (Charlton et al., 2020; Browne 2019). Gibbs sampling is used to choose a starting value for each parameter, and in MLwiN 3.02 (Charlton et al., 2020; Browne 2019) these are identified by running IGLS before running MCMC estimation. According to the recommendations of Draper (2008), a burn-in of 500 simulations was employed for each model, followed by 50,000 monitoring simulations. The convergence of Markov chains was assessed by a visual inspection of the trace- and autocorrelation-plots alongside MCMC diagnostics. These diagnostics include: the Raftery-Lewis diagnostic (Raftery and Lewis 1992), the Brooks-Draper Diagnostic (Brooks and Draper 1999), the Monte-Carlo Standard Error (MCSE) and the effective sample size (ESS) (Kass et al., 1998). Each model was then built, gradually adding in the additional level and predictors one-by-one. The DIC, changes in the estimates of variance and the regression (fixed) coefficients were used to evaluate the goodness-of-fit of each model. The MCMC posterior means and 95% credible intervals are presented as summaries of the posterior distribution of each variable.

The model assumptions were checked using normal probability plots, where ranked residuals are plotted against corresponding points on a Normal distribution curve at each level of the model (Fig. 7). For these plots, if the normality assumption is valid the points should lie on approximately a straight line. These plots also allow the identification of any outliers.

### 3.1. *Detecting variability in continuous data: the small stone spheres assemblage from Akrotiri, Santorini (Thera), Greece*

The site of Akrotiri on the modern-day island of Santorini (Thera) is a well-known Bronze Age town that was destroyed by the eruption of the island's volcano during the Middle Bronze Age (possibly the early 16th century BCE) (about Akrotiri see Doumas 1983) (Fig. 8). In Akrotiri, among a wealth of finds, 746 stone spheres have been catalogued, of which 65% were brought to light in the recent excavations (Valassi forthcoming) (Fig. 9). To date, no similar material has been published from any other Aegean Bronze Age site, which has resulted in different interpretations by different researchers. Marinatos (1971: 28) interpreted the spheres as either sling stones or as tossing balls. Later, this interpretation was rejected by Valassi (forthcoming) and Tzachili (2007). They suggest that the spheres are unlikely to have been sling stones as all other examples from this period, and from later periods, are generally heavier than the majority of the spheres from Akrotiri and are

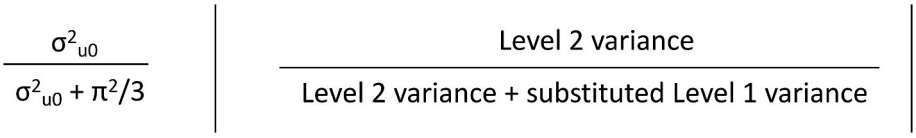


**Fig. 6.** Latent variable approach for the calculation of the VPC for binary logistic response models: formula (left) and general explanation (right) (substituted level 1 variance = 3.29).

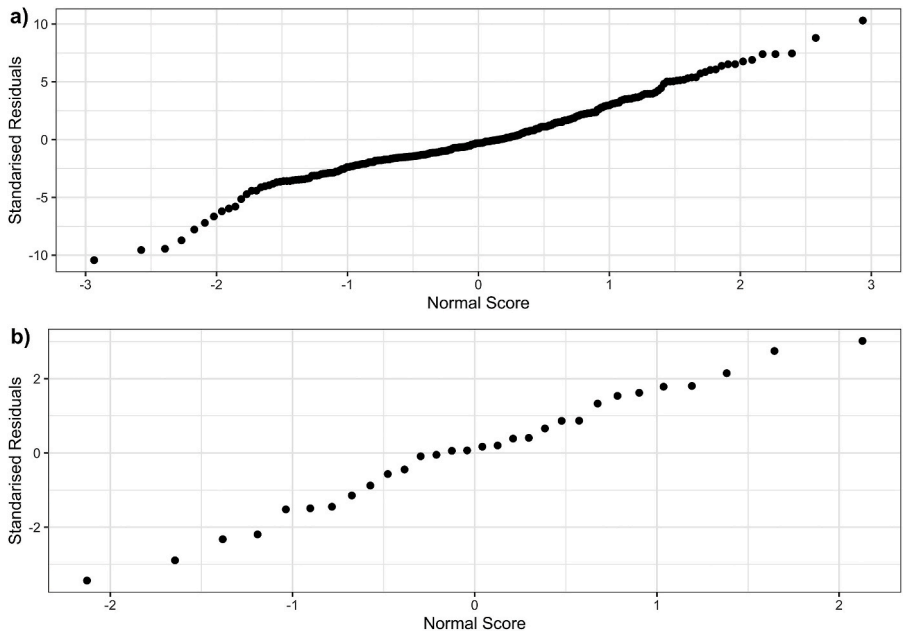


**Fig. 7.** Example normal Plots for a 2-level multilevel model: a) Level 1 and b) Level 2. Ranked residuals are plotted against corresponding points on a Normal distribution curve. These plots can be used to check the assumption that residuals are normally distributed.


**Fig. 8.** A map of the excavated areas of Akrotiri town. The major excavation entities, buildings and open spaces are annotated. The largest groups of worked stone spheres have been unearthed in the entities of Western House, Xeste 3, Sector D, and the Kenotaph square. Map based on (Doumas, 2017).

more ovoid in shape. Valassi and Tzachili suggest they are unlikely to have been used as tossing balls as they could easily harm the players if not caught. However, they agree that the spheres may have been used as a counting/record-keeping system or as "pawns" for a type of board game.

The spheres come in different sizes, colours and stone materials, and have been found throughout the settlement, in both open and closed spaces. MLMs were applied to spheres from the last (Late Cycladic) phase of the settlement to investigate their variability within and between different areas of the settlement. The data has 3 levels - spheres (L1), the excavation entity (buildings' interiors and open spaces with specific boundaries) (L2) and the different building complexes in the town (L3) (Fig. 9). Specifically, this includes 140 spheres, within 19 excavation entities, within 3 zones (Fig. 10). The response variable was sphere diameter, recorded manually using a digital calliper, which ranged between 16 and 57 mm. Sphere diameter was included to test the hypothesis that sphere size was key to its function, being used as a metric tool or as pawns in board games. Two additional sphere level categorical predictor variables were recorded (preservation and worked/natural form). Preservation was included because of its impact on sphere

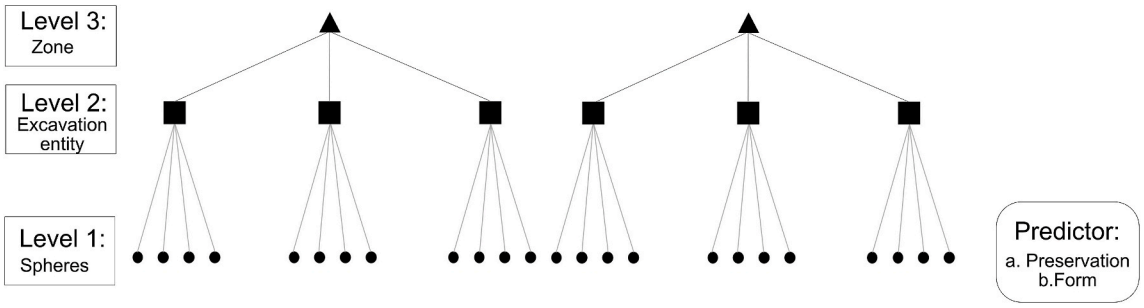


**Fig. 9.** A group of stone spheres from Xeste 3.

diameter. Stone form (worked or natural pebble) was added to investigate the research hypothesis that stones were used for their sphericity and size rather than because they were worked (Fig. 9).

A normal likelihood was fitted to the continuous response data (sphere diameter) with an identity link function. Normal distributions were used for each of the random effects (levels) with an inverse-gamma (0.001, 0.001) prior distribution, and an improper uniform prior of $\propto 1$ was used for the fixed effects (for mathematical structure see supporting information). The sphere data was fitted into 5 models: 1) a null one-level model (C $M_n$), 2) a two-level random-intercept model (spheres and entity) (C $M_2$), 3) a 3-level random-intercept model (sphere, entity, zone) (C $M_3$), 4) a three-level random-intercept model with one L1 predictor (preservation) (C $M_4$), and 5) a three-level random-intercept model with two L1 predictors (preservation and worked) (C $M_5$). The visual inspection of the trace- and autocorrelation-plots as well as the MCMC diagnostics indicated Markov Chain convergence and the normality plots indicated that the residuals at each level were roughly normally distributed (see supporting info for diagnostic statistics and plots and normality plots).

The model summaries are included in Table 1, indicating the posterior distributions and deviance statistics of each model. The change in the DIC across all models is negligible (<2), suggesting that all the models fit the data. When the variance parameters and regression coefficients are interrogated the inclusion of sphere form ("worked") appears to explain substantial variability within zones and the estimated coefficient is large (Table 1). This indicates a model including sphere form but not preservation should be used, the values for this model are provided in Table 3. This also indicates that there is considerable variability in sphere size according to sphere form, with naturally formed stones larger than worked. This is appears to be consistent across context and zones (Fig. 11). The VPC estimates in Table 3 indicate that by far the greatest degree of variation in sphere diameter occurs at the sphere level (L1). This is also evident in the residual plots (Fig. 12). This is followed by, although low, site zone and then entity. For the fixed part of the model, the intercept or sphere mean diameter ranged between 31 and 32 mm depending on the model used.

### 3.2. *Detecting variability in binary data: the pottery assemblage of Mala (nova) Pećina cave, Croatia*

Mala (Nova) Pećina cave is located in the hinterland Dalmatia, Croatia. The cave is deep in the hills overlooking the valley passages that lead from the Herzegovina uplands to the Adriatic coast. Excavations in 2016 discovered evidence of Early and Late Neolithic occupation in the cave (see Trimmis and Drnić 2018; Drnić et al., 2018). Early Neolithic (EN) pottery was mainly concentrated in trench B, which was situated deep in the cave at the end of a long and low passage leading to the third chamber. Late and Early Neolithic pottery were unearthed in trenches A and C, along postholes and hearths in trench A (Trimmis and Drnić 2018:3–4) (Fig. 13).

MLMs were applied to the pottery assemblage from Mala Pećina to identify the potential presence of different activity areas per period in the cave. This data set has three levels: sherd (L1), context (L2) and Trench (L3), with 165 sherds within 25 contexts, within 3 trenches (Fig. 14). The binary response variable in this model was waretype (fine or coarse). A sherd level binary predictor was included, indicating decoration (absence/presence of impressed decoration). These response and predictor variables were included as different ware types are associated with use and, in the case of the impressed ware, are an indication of period. Coarse wear with impressed decoration is an indication of an EN phase (late 7th millennium – middle 6th millennium) Adriatic Neolithic contexts (see Forenbaher et al., 2013: 597).

A binomial likelihood was fitted to the binary response data (waretype) with a logistic link function. Normal distributions were used for each of the random effects (levels) with an inverse-gamma (0.001, 0.001) prior distribution, and an improper uniform prior of $\propto 1$ was used for the fixed effects (for mathematical structure see supporting information). The binary pottery data was fitted to four logistic models: 1) a null one-level model (B $M_n$), 2) a two-level random-intercept model (sherd and context) (B $M_2$), 3) a three-level random-intercept model (sherd, context and trench) (B $M_3$), and 4) a three-level random-intercept model with one level 1 predictor (decoration: impressed or no impressed decoration) (B $M_4$). An inspection of the trace- and autocorrelation-plots indicated chain convergence. However, the MCMC diagnostics (particularly the Raftery-Lewis, Brooks-Draper and ESS) that the chain may not have been run for long enough (see supporting info for diagnostic statistics).

The model summaries are included in Table 4, including the posterior mean, 95% credible interval and the DIC. There is a considerable reduction in the DIC from the $BM_n$ to B $M_3$. This reduces fractionally with the addition of the predictor variable, decoration, however it has a substantial regression coefficient (Table 4).

The VPC estimates presented in Table 5 and the residuals in Fig. 15a and b indicate that sherd waretype varies more between trenches than between contexts. The specific variation in waretype can be attributed from Fig. 15b, trench 1 (A) has a greater log-odds of fine ware compared to trench 2 which has a greater log-odds of coarseware.

For the fixed part of the model, the intercept of the logistic regression reflects the log-odds of waretype. According to the B $M_2$ intercept in Table 4, the log-odds of a sherd being fine ware is less likely. However, in B $M_4$ the log-odds is more likely for a sherd to be fine ware with no decoration (Table 4). The decoration regression coefficient also indicates a considerable amount of variability in waretype according to decoration (Table 4). This variation can be seen between both contexts and trenches in Fig. 15c and d, with fine ware having a greater odds of being non-impressed and coarse ware a greater odds of being impressed. At the trench level, in trench 2 (B) there is greater odds of sherds being coarse with impresso decoration (Fig. 15d).

## 4. Discussion

MLMs allow the inclusion of explanatory variables and different levels to explain variation in the dependent variable. The variability in the random effects can be determined and proportioned using the VPC

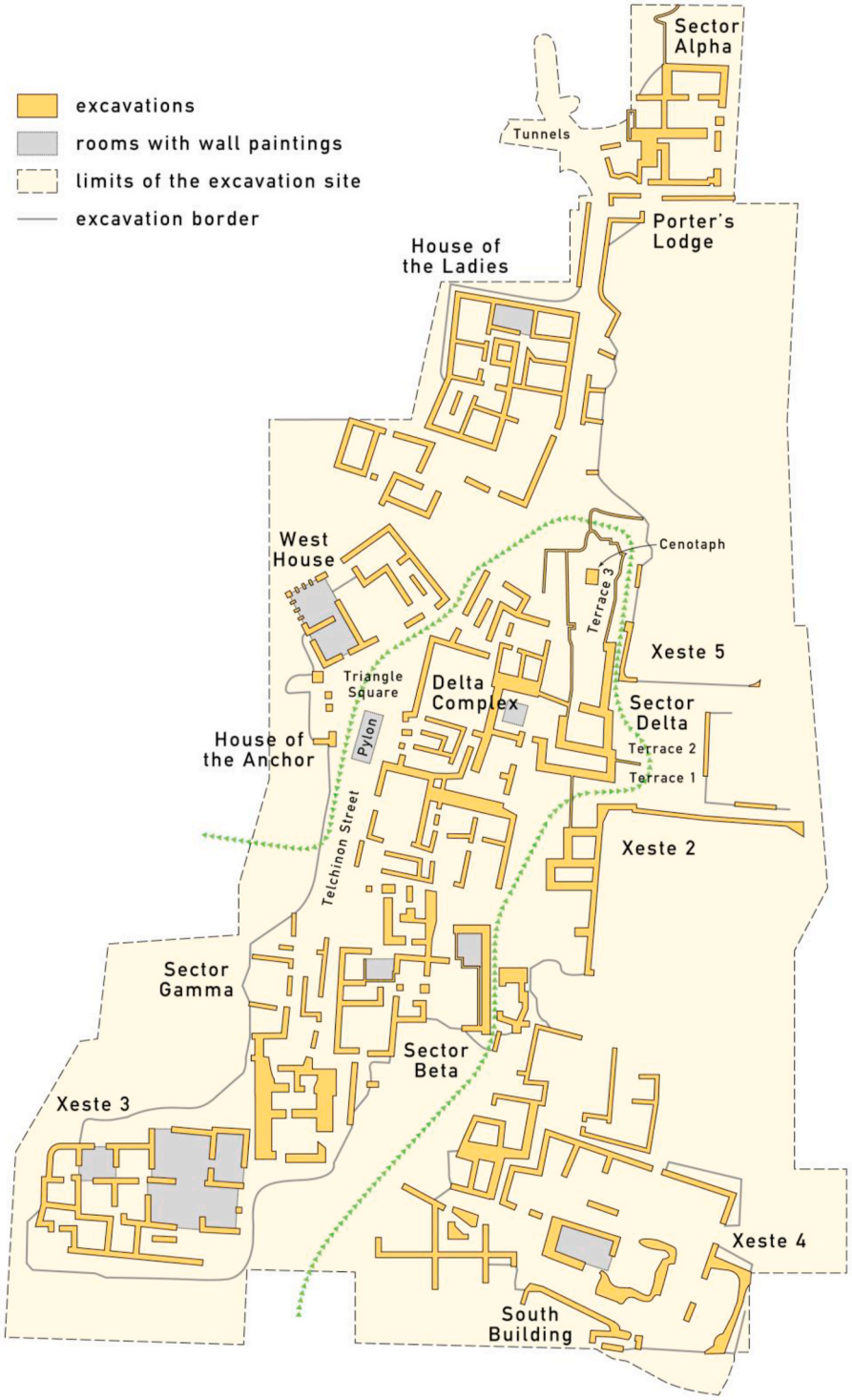


**Fig. 10.** Diagram of Akrotiri sphere multilevel model structure. Speres (circles) are the level 1 units grouped into excavation entity (squares) at level 2, within zones (triangles) level 3. The model also includes fixed explanatory predictor variables: sphere preservation and form at level 1.

(given in the random part of Tables 1, 3 and 4). Although this paper largely focuses on the random effects, the contribution and the variability of the explanatory variables, or fixed effects, can also be interpreted through changes in the fixed regression coefficients and the variance parameters. The fixed regression coefficients (given in the fixed part of Tables 1, 3 and 4) are indicative of variability in the data. A higher coefficient mean value indicates a greater degree of variation. Finally, the intercept can be used to indicate the overall mean of the dependent variable, sphere diameter in the Akrotiri model and waretype in the Mala Pećina model.

The application of MLMs to both assemblage's aid in their interpretation. The Akrotiri stone sphere MLMs indicate greater variability at the sphere level compared to excavation entity. This can be interpreted as spheres clustering in groups of different sizes within the different excavation entities and zones of the excavated town. This accords with previously published research on the Aktoriri sphere assemblage. For example, a group of spheres excavated from the Western House are a variety of sizes, from very small (16 mm) to large (45 mm). It was from

**Table 1**
Akrotiri Sphere MLM MCMC posterior means (mm) with the 95% central interval (Bayesian credible interval) for each parameter in the model. Variance estimates ($\sigma^2$) are given for each of the random variables (sphere, entity and zone). This is accompanied by the DIC, the effective number of parameters (pD) and a deviance statistic evaluating the posterior mean of the model parameters (Dthebar) for each model.

| Parameter | C $M_n$ | C $M_2$ | C $M_3$ | C $M_4$ | C $M_5$ |
|---|---|---|---|---|---|
| Random | | | | | |
| Sphere | 50.821 (40.138, 64.423) | 50.321 (39.502, 63.646) | 50.328 (39.590, 63.362) | 50.626 (50.155, 64.438) | 49.930 (39.174, 63.401) |
| Entity | | 1.00 (0.001, 7.536) | 1.279 (0.001, 8.798) | 1.373 (0.001, 9.414) | 1.250 (0.001, 8.755) |
| Zone | | | 5.622 (0.001, 16.013) | 6.306 (0.001, 17.252) | 2.048 (0.001, 9.909) |
| Fixed | | | | | |
| Intercept | 32.178 (30.995, 33.361) | 32.178 (30.855, 33.505) | 32.282 (30.434, 34.207) | 32.233 (32.145, 34.192) | 31.516 (29.621, 33.419) |
| Preservation | | | | 0.688 (−3.568, 4.967) | 1.257 (−3.044, 5.497) |
| Worked | | | | | 2.652 (−0.333, 5.627) |
| DIC | 948.20 | 948.41 | 948.97 | 950.87 | 949.73 |
| pD | 1.985 | 3.513 | 4.222 | 5.312 | 6.130 |
| D (thebar) | 944.23 | 941.39 | 940.52 | 945.56 | 937.45 |
| Dbar | 946.213 | 944.899 | 944.744 | 942.984 | 943.62 |

**Table 2**
VPC estimate (%) for levels (sphere (L1), entity (L2), and zone (L3)) in each model.

| Parameter | C $M_n$ | C $M_2$ | C $M_3$ | C $M_4$ | C $M_5$ |
|---|---|---|---|---|---|
| Sphere | – | 98.05 | 87.94 | 86.84 | 93.80 |
| Entity | – | 1.95 | 2.24 | 2.35 | 2.35 |
| Zone | – | – | 9.82 | 10.81 | 3.85 |

**Table 3**
Model estimates for Akrotiri Sphere three-level model with one Level 1 predictor (worked). MCMC posterior means (mm) with the 95% central interval (Bayesian credible interval) for each parameter in the model. Variance estimates ($\sigma^2$) are given for each of the random variables (sphere, entity and zone). This is accompanied by the DIC and the effective number of parameters (pD) and a deviance statistic evaluating the posterior mean of the model parameters (Dthebar) for each model.

| Random | Model Estimates | VPC |
|---|---|---|
| Parameter | | |
| Sphere | 49.700 (39.083, 63.193) | 89.01 |
| Entity | 1.184 (0.001, 8.262) | 2.12 |
| Zone | 4.951 (0.001, 13.797) | 8.87 |
| Fixed | | |
| Intercept | 31.722 (29.783, 33.663) | |
| Worked | 2.505 (−0.429, 5.433) | |
| DIC | 948.07 | |
| pD | 5.08 | |
| D (thebar) | 937.90 | |
| Dbar | 942.984 | |

this context that the hypothesis of the spheres being used as a counting system or game board 'pawns' were born (Tzachili 2007). The use of MLMs has elevated a theory based on small excavated groups of spheres

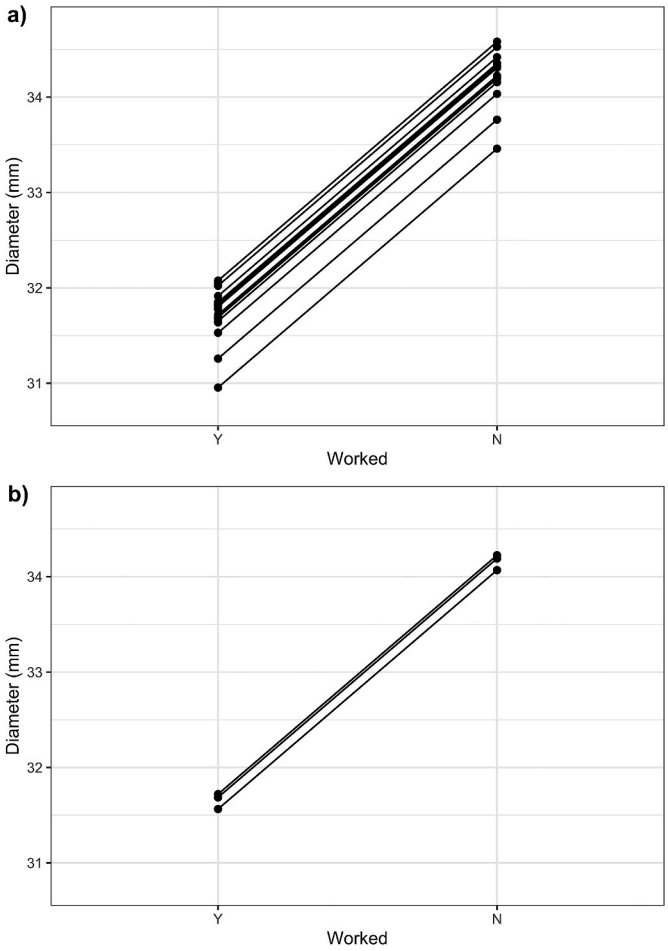


**Fig. 11.** Akrotiri model group predictions of sphere diameter against sphere form (worked or natural). Group intercepts allowed to vary by excavation unit (a) and zone (b).

into an indication of patterns, or lack thereof, occurring across the excavated town. In addition to this, interpretations can be made regarding spheres preservation and form. The higher regression coefficient of sphere form in comparison to preservation, in Tables 2 and 3, suggests that worked spheres were on an average, larger than natural spheres. However, the diminutive impact of the sphere level predictor on the VPC estimates, in Table 3, supports the hypothesis that stone form, natural or worked, was chosen indiscriminately. Spheres appear to have been worked or selected to fit certain dimensions.

The MLM results for the pottery assemblage from Mala Pećina confirmed the hypothesis that pottery varies, in both waretype and decoration, between trenches. The variability in decoration waretype regression coefficients is consistent with characteristic patterns of decoration and ware observed in EN contexts. Trench B (1) showed a unified group of EN impressed pottery, compared to trench A and C (2 and 3) where pottery evidence was mixed with EN pottery in the lower strata and LN pottery in the upper layers (see also Drnić et al., 2018). Testing this outcome against the excavation evidence, occupation in Trench B during EN seems more intense, with high quantities of pottery in a thin stratigraphic palimpsest (Trimmis and Drnić 2018). Greater variability when impressed decoration is included as a predictor, supports the excavator hypothesis that not all areas of the cave were used simultaneously, as the Trench A shows mainly LN pottery. The increase in waretype variation showcases the phenomenon that EN impressed pottery is generally represented on coarse rather than fine waretypes (see Table 4). The variability of pottery waretype remains high in the model with no decoration predictor (B $M_3$), this indicates that particular

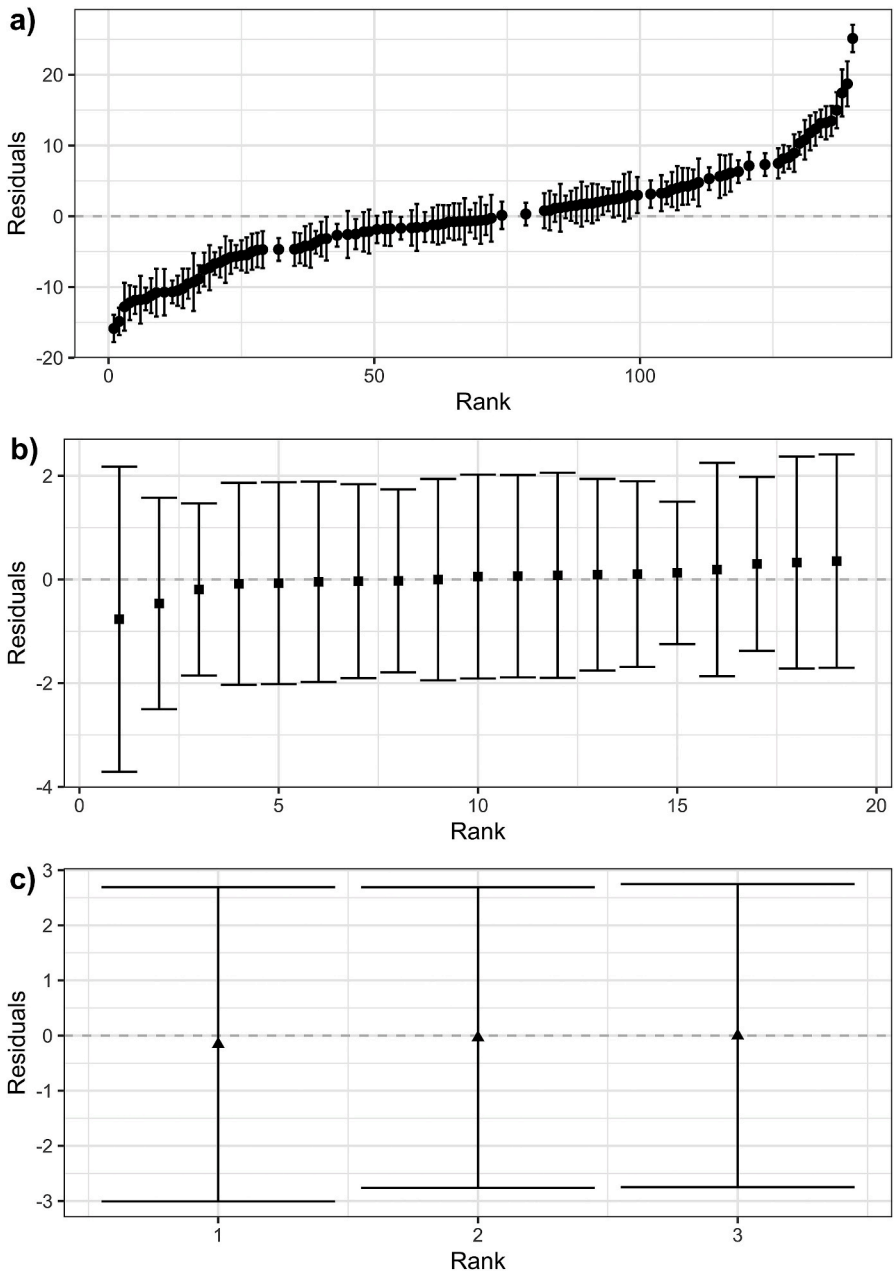


**Fig. 12.** Residual plots for Akrotiri multilevel model: a) Level 1 (Spheres), b) Level 2 (Excavation units) and c) Level 3 (Zones). Residuals are plotted for each 'group' (Level 1: Sphere, Level 2: Excavation unit and Level 3: Trench) accompanied by their confidence intervals. The overall average (the fixed parameter $\beta_0$) is represented by the dashed line.

areas of the cave are concentrated with certain waretypes, something that corresponds well with the excavation observations.

The models that are presented in this paper were both single site pilot studies of the applicability and functionality of multilevel modelling for the analysis of archaeological data. In both cases, the datasets were small and model outcomes could be easily evaluated based on assumptions from previous excavators at each site. The MLM results, in both cases, support the initial hypotheses. In Akrotiri, the results suggest that the use of these spheres may not have differed between the site zones and entities, which opens up a new avenue of research on their potential function.

With the incorporation of additional data, there is potential for the addition of further levels to both sets of models. A fourth level, 'cave', and a fifth level, 'region' could be added to the Mala Pećina pottery models. This would facilitate the investigation of regional variations in Neolithic pottery, for example variations between caves from the coast, hinterland and uplands. Site and Island (Cycladic) levels could also be added to the Akrotiri model. This would enable the analysis of variability in lithic spheres across the Cyclades and contribute to the discussion about the spheres' function in the Bronze age Aegean.

The Akrotiri sphere MLM was modelled using a Normal likelihood and identity link function which is commonly used for continuous data. However, using this link function allows diameter to take negative values. As diameter cannot be negative an alternative link function, such as a log link function, could be used to account for this. This link can be used when a value is constrained to be positive. However, if this function is used the VPC must be estimated using the methods for binary response models such as those used here for the Mala Pećina pottery data.

The MCMC diagnostics and plots for the Mala Pećina indicate issues with convergence of and autocorrelation in the MCMC chain (supporting

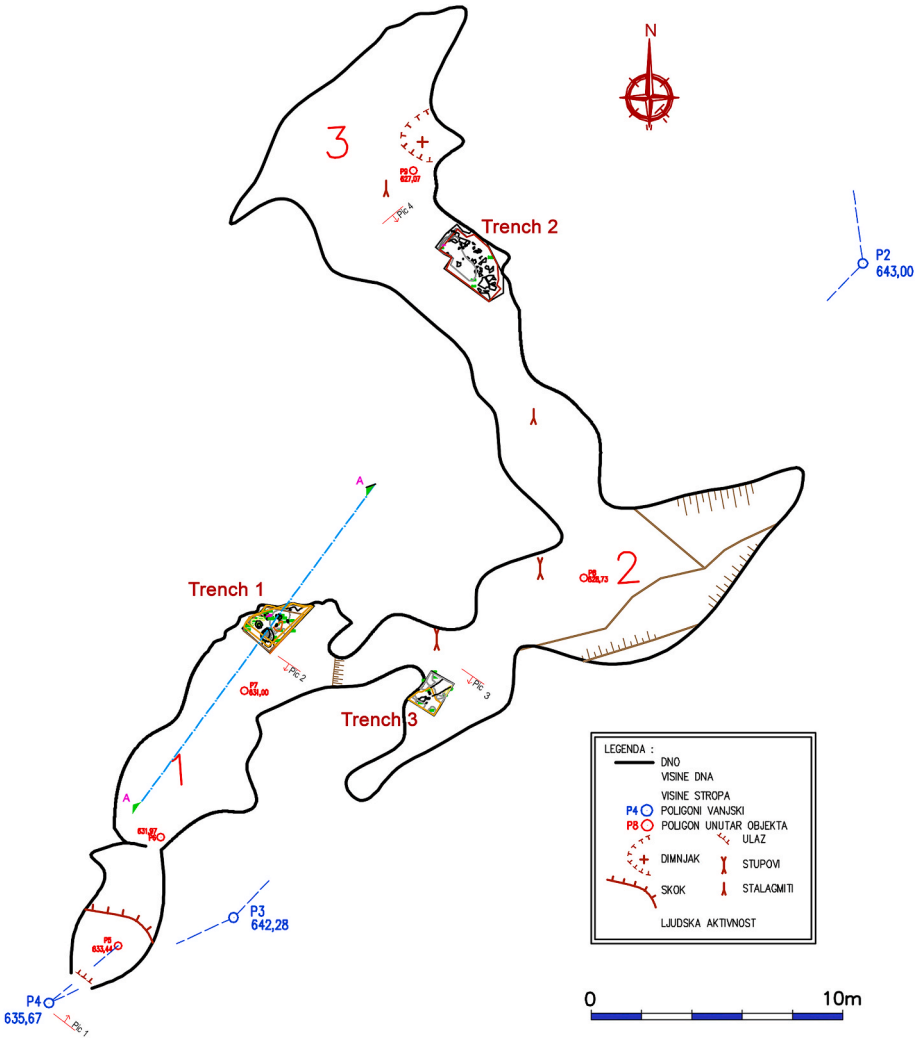


**Fig. 13.** A ground plan of Mala (Nova) Pećina. The three excavations trenches are annotated. Map adapted from Drnić et al. (2018).

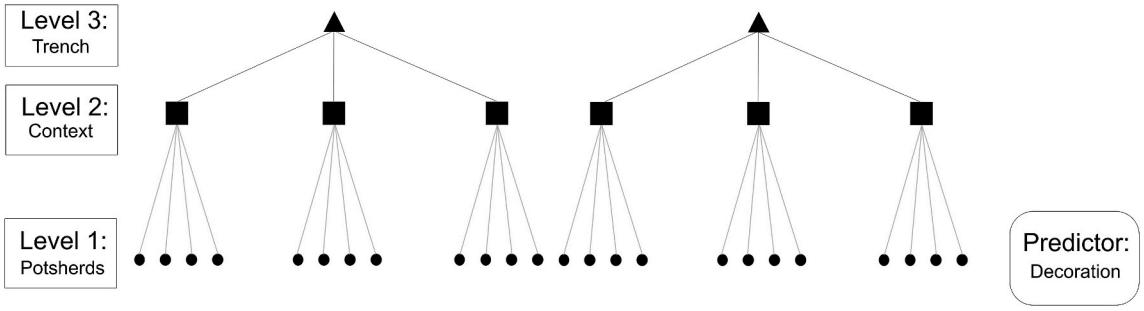


**Fig. 14.** Diagram of Mala Pećina multilevel model structure. Pottery sherds (circles) are the level 1 units grouped into contexts (squares) at level 2, within trench (triangles) level 3. The model also includes a fixed explanatory predictor variables: sherd decoration at level 1.

information). This suggests that the chain was not run for long enough. This may result in issues with the posterior distributions, specifically the mean and 95% credible intervals, estimated for the model parameters. To overcome this issue the chain length can be increased to more than 50,000 iterations. Thinning the MCMC chain is another possibility, this results in the discarding of all but every kth sampled value in the chain. However, the use of thinning to overcome autocorrelation has been debated (Link and Eaton 2012).

This paper has focused largely on the variability occurring in the random part of the respective models, across the levels. Further analysis can be carried out on the fixed predictive variables (see Fernée 2020), which can identify differences between the predictive variables and their specific contribution to the model. For the binary model, it is also possible to calculate probabilities of the outcome variable based on the fixed parameters (see Austin and Merlo, 2017).

Further studies are needed to confirm further possible applications of MLMs in archaeology. The two pilot case studies confirmed that MLMs can both detect variability and highlight specific patterns occurring within and between levels. Multilevel models may withstand the criticism of the positivism of archaeological statistics due to the elasticity in

**Table 4**
Mala Pećina sherd MLM MCMC posterior means (log odds ratios) with the 95% central interval (Bayesian credible interval) for each parameter in the model. Variance estimates ($\sigma^2$) are given for each of the random variables (sherd, context and trench). This is accompanied by the DIC and the effective number of parameters (pD) and a deviance statistic evaluating the posterior mean of the model parameters (Dthebar) for each model.

| Parameter | B $M_n$ | B $M_2$ | B $M_3$ | B $M_4$ |
|---|---|---|---|---|
| Random | | | | |
| Sherd | – | – | – | – |
| Context | | 1.138 (0.210, 3.599) | 0.956 (0.004, 3.581) | 1.073 (0.008, 3.851) |
| Trench | | | 1.161 (0.001, 6.777) | 2.001 (0.001, 12.969) |
| Fixed | | | | |
| Intercept | −0.234 (−0.548, 0.072) | −0.071 (−0.744, 0.566) | −0.181 (−1.618, 0.815) | 0.208 (−1.961, 1.461) |
| Decoration | | | | −0.848 (−1.630, −0.102) |
| DIC | 228.56 | 206.47 | 208.08 | 204.77 |
| pD | 1.01 | 10.84 | 10.71 | 12.15 |
| D (thetabar) | 226.55 | 184.80 | 186.66 | 180.48 |
| Dbar | 227.55 | 195.63 | 197.37 | 192.62 |

**Table 5**
VPC estimate (%) for levels (Context (L2), and Trench (L3)) in each model.

| Parameter | B $M_n$ | B $M_2$ | B $M_3$ | B $M_4$ |
|---|---|---|---|---|
| Sherd | – | – | – | – |
| Context | – | 25.70 | 17.68 | 16.86 |
| Trench | – | – | 21.47 | 31.44 |

their application, structured nature, and the incorporation of both interpretational predictors and Bayesian statistics. Initial model construction is hypothesis driven. The levels and their number as well as the predictors selected are interpretational, and they can incorporate the archaeologist's perspective into the analysis. In the Akrotiri case study, the model structure was guided by the hypothesis of sphere function by Valassy and Tzachili, with the selection of excavation entity rather than excavation context.

Unlike existing methods used in archaeology, MLMs allow the concurrent analysis of multi-layered archaeological data whilst incorporating the researchers' theoretical ideas into empirical testing. MLMs are a potential tool to bridge the gap between current archaeological theory and statistical analysis, particularly if we understand the identification and interpretation of the relationship between different archaeological

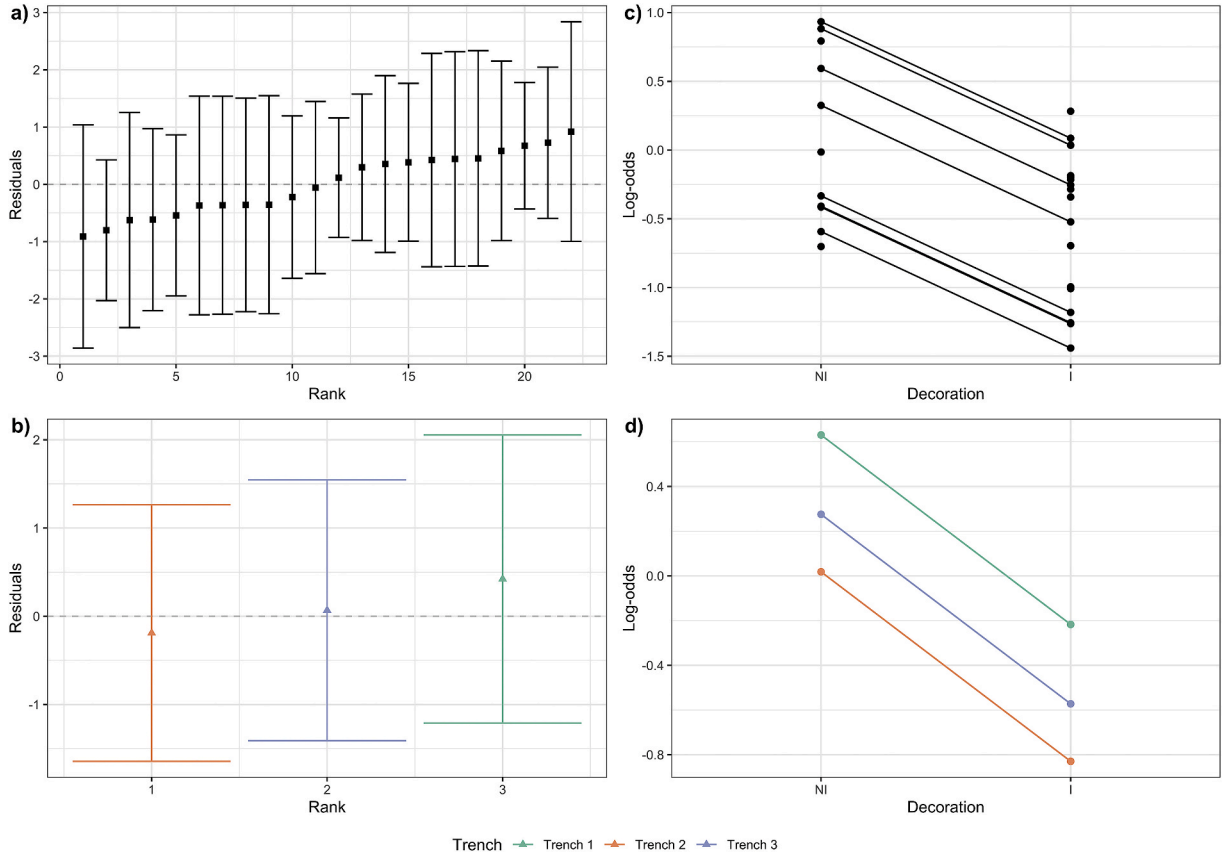


**Fig. 15.** Plots illustrating the random and fixed effects in the Mala Pecina Model. Left: the residual log-odds of the sherds being fine ware are plotted by context (a) and trench (b) accompanied by their confidence intervals. The overall average (the fixed parameter $\beta_0$) is represented by the dashed line Right: the log-odds of sherd ware type against decoration with random intercepts by context (c) and trench (d). Decoration: not impressed (NI) and impressed (I).

evidence and the researcher as the new frontier for the archaeological thought (see Harris and Cipola 2017: 195).

## 5. Conclusion

It is evident from this pilot application that multilevel modelling can be applied to an array of archaeological materials and contexts, from artefact analysis to site pattern interpretation. MLMs can be simple to employ and can be carried out in specific software such as MLwiN 3.05 (Charlton et al., 2020; Browne 2019), and software environments, such as R (R Core Team 2018). In each case study, the MLMs provide indications of the level at which variation occurs, the 'trench' level for the Mala Pećina pottery model, or at the excavation entity/zone level for Akrotiri spheres. They also indicate how the predictor variables contribute to this variation. The main strengths of MLMs, compared to other methods for detecting variability in archaeological data, are the ability to concurrenty study data at multiple levels, the incorporation of predictors as a further interpretational tool and the inclusion of Bayesian statistics.

## Data availability statement

All data supporting this study and the methodology used are openly available from the following repository at 10.6084/m9.figshare.13796228.

## Acknowledgements

The application of multilevel modelling to archaeological data was part of the PhD research by C.L.F funded by AHRC-SWW DTP, UK and British Association of Biological Anthropology and Osteoarchaeology, UK. The relationship between spatial statistics and archaeological theory, and the excavations in Mala Pećina were part of the PhD research by K.P.T supported by Matti Egon II scholarship of the Greek Archaeological Committee, UK and the British Cave Research Association, UK. Research on the Akrotiri spheres has been partially supported by the KA201-079065 EU grant. Authors would like to thank, Dr Ivan Drnić for providing access to the Mala Pećina data, Dr Tania Devetzi and Professor Christos Doumas who made the material from Akrotiri spheres available, and Argyris Mavromatis, Lefteris Zorzos and Maria Karra for their help on Santorini.

## Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jas.2021.105346.

## References

Austin, P.C., Stryhn, H., Leckie, G., Merlo, J., 2017. Measures of clustering and heterogeneity in multilevel Poisson regression analyses of rates/count data. Stat. Med. 37, 572–589.
Banks, W.E., Bertran, P., Ducasse, S., Klaric, L., Lanos, P., Renard, C., Mesa, M., 2019. An application of hierarchical Bayesian modeling to better constrain the chronologies of Upper Paleolithic archaeological cultures in France between ca. 32,000-21,000 calibrated years before present. Quat. Sci. Rev. 220, 188–214.
Bates, D., Mächler, M., Bolker, B., Walker, S., 2015. Fitting linear mixed-effects models using lme4. J. Stat. Software 67 (1), 1–48. https://doi.org/10.18637/jss.v067.i01.
Brooks, S.P., Draper, D., 1999. Comparing the efficiency of MCMC samplers. Technical report. In: Department of Mathematical Sciences. University of Bath, UK.
Browne, W.J., 2019. MCMC estimation in MLwiN. In: Bristol: Centre for Multilevel Modelling. University of Bristol, 3.03.
Browne, W.J., Draper, D., 2006. A comparison of Bayesian and likelihood-based methods for fitting multilevel models. Bayesian Analysis. 1 (3), 473–514.
Browne, W.J., Subramanian, S.V., Jones, K., Goldstein, H., 2005. Variance partitioning in multilevel logistics models with over-dispersion. J. Roy. Stat. Soc. 168 (3), 599–613. https://doi.org/10.1111/j.1467-985X.2004.00365.x.
Brunton-Smith, I., Sturgis, P., 2011. Do neighborhoods generate fear of crime? An empirical test using the British crime survey. Criminology. https://doi.org/10.1111/j.1745-9125.2011.00228.x, 49, 2, 331–369.
Bürkner, P., 2017. Brms: an R package for bayesian multilevel models using stan. J. Stat. Software 80 (1), 1–28. https://doi.org/10.18637/jss.v080.i01.
Charlton, C., Rasbash, J., Browne, W.J., Healy, M., Cameron, B., 2020. MLwiN. In: Centre for Multilevel Modelling. University of Bristol, Version 3.05.
Dias, S., Sutton, A.J., Welton, N.J., Hall, C., Road, W., Unit, D.S., Court, R., 2011. NICE DSU Technical Support Document 3: Heterogeneity: Subgroups, Meta-Regression, Bias and Bias-Adjustment. Report by the Decision Support Unit.
Doran, J.E., Hodson, F.R., 1975. Mathematics and Computers in Archaeology. Harvard University Press, Cambridge.
Doumas, C., 1983. Thera - Scavi a Santorini. Thames and Hudson, London.
Doumas, C, 2017. Akrotiri the Archaeological Site and the Museum of Prehistoric Thera. Society for the promotion of studies on prehistoric Thera, Athens.
Draper, D., 2008. Bayesian Multilevel Analysis and MCMC. In: Deleeuw, J., Meijer, E. (Eds.), Handbook of Multilevel Analysis. Springer, New York, pp. 77–139.
Drennan, R.D., 2009. Statistics for Archaeologists. In: Interdisciplinary Contributions to Archaeology, second ed. Springer US, Boston, MA http://link.springer.com/10.1007/978-1-4419-0413-3.
Drnić, I., Trimmis, K., Hale, A., Madgwick, R., Reed, K., Barbir, A., Mađerić, M., 2018. Assemblages from marginal spaces: the results of the excavations in Mala (nova) Pećina near Muć and the neolithic of dalmatinska zagora. Prilozi instituta za arheologiju u zagrebu 32: 29 – 70. http://hdl.handle.net/1983/38b0cc7d-71bc-42c5-81eb-1b389b3866e3.
Earle, T.K., Preucel, R.W., 1987. Processual archaeology and the radical critique. Curr. Anthropol. 28 (4), 501–538. http://www.jstor.org/stable/2743487.
El-Horbaty, Y.S., Hanafy, E.M., 2018. Some estimation methods and their assessment in multilevel models: a review. Biostatistics and Biometrics Open Access Journal 5 (3), 1–8. https://doi.org/10.19080/BBOAJ.2018.04.555662.
Fernée, C., 2020. Like Pulling Teeth: A Study of Variation in Tooth Size and Shape in Historic and Modern Populations. PhD Thesis. University of Southampton.
Forenbaher, S., Kaiser, T., Miracle, P.T., 2013. Dating the east Adriatic Neolithic. Eur. J. Archaeol. 16 (4), 589–609. https://doi.org/10.1179/1461957113Y.0000000038.
Garvey, R., 2018. Current and potential roles of archaeology in the development of cultural evolutionary theory. Phil. Trans. R. Soc. B 373, 20170057. https://doi.org/10.1098/rstb.2017.0057.
Gayo, E.M., Latorre, C., Santoro, C.M., 2015. Timing of occupation and regional settlement patterns revealed by time-series analyses of an archaeological radiocarbon database for the South-Central Andes (16°–25°S). Quarernary International. 356, 4–14. https://doi.org/10.1016/j.quaint.2014.09.076.
Gelman, A., 2006. Multilevel (hierarchical) modeling: what it can and cannot do. Technometrics 48, 432–435. https://doi.org/10.1198/004017005000000661.
Gelman, A., Hill, J., 2007. Data Analysis Using Regression and Multilevel/Hierachical Models. Cambridge University Press, Cambridge.
Gelman, A., Simpson, D., 2017. The prior can often only be understood in the context of the likelihood. Entropy 19 (10), 555. https://doi.org/10.3390/e19100555.
Gelman, A., Carlin, J.B., Stern, H.S., Dunson, D.B., Vehtari, A., Rubin, D.B., 2013. Bayesian Data Analysis, third ed. CRC Press, , Boca Raton.
Goldstein, H., 1995. Multilevel Statistical Models. Edward Arnold, London.
Goldstein, H., 2002. Multilevel models where the random effects are. Group.
Goldstein, H., 2010. Multilevel Statistcal Models 4th Ed. John Wiley & Sons, London.
Goldstein, H., Browne, W., Rasbash, J., 2002. Partitioning variation in multilevel models. Understand. Stat. 1 (4), 223–231. https://doi.org/10.1207/S15328031US0104_02.
Harris, O.J.T., Cipolla, G.N., 2017. Archaeological Theory in the New Millennium. Routledge, London and New York.
Hodder, I., 1986. Reading the Past: Current Approaches to Interpretation in Archaeology. Cambridge University Press, Cambridge.
Hole, B.L., 1980. Sampling in archaeology: a critique. Annu. Rev. Anthropol. 9, 217–234. https://doi.org/10.1146/annurev.an.09.100180.001245.
Hox, J.J., 2010. In: Multilevel Analysis: Techniques and Applications, 2nd. Routledge, New York.
Hurst Thomas, D., 1978. The awful truth about statistics in archaeology. Contributions to archaeological method and theory. Am. Antiq. 43 (2), 231–244. https://doi.org/10.2307/279247.
Jones, K., Subrmanian, S.V., 2013. Developing Multilevel Models for Analysing Contextuality, Heterogeneity and Change Using MLwiN 2.2. Centre for Multilevel Modelling. University of Bristol.
Jones, K., Subrmanian, S.V., 2017. Developing Multilevel Models for Analysing Contextuality, Heterogeneity and Change Using MLwiN 3.0. Centre for Multilevel Modelling. University of Bristol.
Julian, M.W., 2001. The consequences of ignoring multilevel data structures in nonhierarchical covariance modeling. Struct. Equ. Model. 8, 325–352. https://doi.org/10.1207/S15328007SEM0803_1.
Kass, R.E., Carlin, B.P., Gelman, A., Neal, R.M., 1998. Statistical practice. Am. Statistician 52 (2), 93–100.
Kent, S., 1992. Studying variability in the archaeological record: an ethnoarchaeological model for distinguishing mobility patterns. Am. Antiq. 57 (4), 635–660. https://doi.org/10.2307/280827.
Kharazifard, M.J., Holakouie-Naieni, K., Mansournia, M.A., 2017. Application of multilevel models in dentistry. J. Dent. Tehran Univ. Med. Sci. 14 (6), 352–360.
Kim, J., Marcusson-Clavertz, D., Togo, F., Park, H., 2018. A practical guide to analyzing time-varying associations between physical activity and affect using multilevel modeling. Computational and Mathematical Methods in Medicine 2018, 1–11. https://doi.org/10.1155/2018/8652034.
Leckie, G., Merlo, J., Austin, P., 2019. Variance partitioning in multilevel models for count data. Available at: arXiv:1911.06888v1. (Accessed 21 January 2020).
Link, W.A., Eaton, M., 2012. On thinning of chains in MCMC. Methods in Ecology and Evolution. 2, 112–115. https://doi.org/10.1111/j.2041-210X.2011.00131.x.

Liu, C., Shimelmitz, R., Friesem, D.E., Yeshurun, R., Nadel, D., 2020. Diachronic trends in occupation intensity of the Epipaleolithic site of Neve David (Mount Carmel, Israel): a lithic perspective. J. Anthropol. Archaeol. 60, 101233. https://doi.org/10.1016/j.jaa.2020.101223.

Maas, C.J.M., Hox, J.J., 2004. The influence of violations of assumptions on multilevel parameter estimates and their standard errors. Comput. Stat. Data Anal. 46, 427–440. https://doi.org/10.1016/j.csda.2003.08.006.

Marinatos, S., 1971. Excavation at Thera V. Athens.

McCorriston, J., 2002. Spatial and temporal variation in mesopotamian agricultural practices in the khabur basin, Syrian jazira. J. Archaeol. Sci. 29, 485–498. https://doi.org/10.1006/jasc.2001.0741.

Orton, C., 1980. Mathematics in Archaeology. Cambridge University Press.

Otárola-Castillo, E., Torquato, M.G., 2018. Bayesian statistics in archaeology. Annu. Rev. Anthropol. 47 (1), 435–453.

O'Malley, M.A., Brigandt, I., Love, A.C., Crawford, J.W., Gilbert, J.A., Knight, R., Mitchell, S.D., Rohwer, F., 2014. Multilevel research strategies and biological systems. Philos. Sci. 81 (5), 811–828. https://doi.org/10.1086/677889.

O'Shea, J.M., 1984. Mortuary Variability. An Archaeological Investigation. Academic Press, Orlando FL.

Perri, A.R., Koster, J.M., Otárola-Castillo, E., Burns, J.L., Cooper, C.G., 2019. Dietary variation among indigenous Nicaraguan horticulturalists and their dogs: an ethnoarchaeological application of the Canine Surrogacy Approach. J. Anthropol. Archaeol. 55, 101066. https://doi.org/10.1016/j.jaa.2019.05.002.

Plog, S., 1980. Stylistic Variation in Prehistoric Ceramics: Design Analysis in the American Southwest. Cambridge University Press, Cambridge.

Plummer, M., Stukalov, A., Denwood, M., 2019. Rjags: bayesian graphical models using MCMC.

R Core Team, 2018. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Raftery, A.E., Lewis, S.M., 1992. How many iterations in the Gibbs sampler?. In: Bernardo, J.M., Smith, A.F.M., Dawid, A.P., Berger, J.O. (Eds.), Bayesian Statistics, vol. 4. Oxford University Press, Oxford, pp. 763–773.

Rasbash, J., Steele, F., Browne, W.J., Goldstein, H., 2017. A User's Guide to MLwiN. Centre for Multilevel Modelling, University of Bristol, Bristol.

Roberts, B.W., Vander Linden, M., 2011. Investigating Archaeological Cultures: Material Culture, Variability, and Transmission. Springer, London. https://doi.org/10.1007/978-1-4419-6970-5_1.

Saltelli, A., 2008. What is sensitivity analysis. In: Saltelli, A., Chan, K., Scott, E.M. (Eds.), Sensitivity Analysis. John Wiley & Sons, Chichester, pp. 3–12.

Schiffer, M.B., Skibo, J.M., 1997. The explanation of artifact variability. Am. Antiq. 62, 27–50. https://doi.org/10.2307/282378.

Schmid, C., 2019. Evaluating cultural transmission in Bronze age burial rites of central, northern and northwestern europe using radiocarbon data. Adapt. Behav. 28, 359–376. https://doi.org/10.1177/1059712319860842.

Shanks, M., Tilley, C., 1982. Ideology, symbolic power, and ritual communication: a reinterpretation of Neolithic mortuary practices. In: Hodder, I. (Ed.), Symbolic and Structural Archaeology1. Cambridge University Press, Cambridge, pp. 29–154.

Spiegelhalter, D.J., Best, N.G., Carlin, B.P., van der Linde, A., 2002. Bayesian measures of model complexity and fit. J. Roy. Stat. Soc. B 64 (4), 583–639. https://doi.org/10.1111/1467-9868.00353.

Stan Development Team, 2020a. CmdStanR. http://mc-stan.org/ a, R packageversion 0.3.0.

Stan Development Team, 2020b. RStan: the R interface to Stan. http://mc-stan.org/ b, R packageversion 2.21.2.

Steenbergen, M.R., Jones, B.S., 2002. Modeling multilevel data structures. Am. J. Polit. Sci. 46 (1), 218–237. https://doi.org/10.2307/3088424.

Sullivan III, A.P., Olszewski, D.I., 2016. Archaeological Variability and Interpretation in Global Perspective. University Press of Colorado, Boulder.

Theobald, E., 2018. Students are rarely independent: when, why, and how to use random effects in discipline-based education research. CBE-Life Sci. Educ. 17 (3), rm2. https://doi.org/10.1187/cbe.17-12-0280.

Trimmis, K.P., Drnić, I., 2018. Connecting early neolithic worlds: excavating Mala (nova) Pećina in dalmatian zagora, Croatia. Antiquity 92 (362). https://doi.org/10.15184/aqy.2018.57.

Tzachili, I., 2007. Poikila. In: Doumas, Ch (Ed.), Akrotiri Thera, Western House. Archaeological Society at Athens, Athens, 256 – 158.

Valacy, L. (forthcoming) the Small Stone Spheres from Akrotiri, Thera. In Doumas, Ch. (ed) Akrotiri, 40 Years of Research. Athens: Archaeological Society of Athens,.

White, J.P., Hurst Thomas, D., 1972. What mean these stones? In: Clark, D.L. (Ed.), Ethno-taxonomic Models and Archaeological Interpretations in the New Guinea Highlands. Models in archaeology, London: Methuen, pp. 275–308.

Wolfhagen, J., 2020. Re-examining the use of the LSI technique in zooarchaeology. J. Archaeol. Sci. 123, 105254. https://doi.org/10.1016/j.jas.2020.105254.

Zhang, Z., Parker, R.M.A., Charlton, C.M.J., Leckie, G., Browne, W.J., 2016. R2MLwiN: a package to run MLwiN from within R. J. Stat. Software 72 (10), 1–43. https://doi.org/10.18637/jss.v072.i10.