A statistical sub-sampling tool for extracting vegetation community and diversity information from pollen assemblage data

Keen, Hayley F.; Gosling, William D.; Hanke, Felix; Miller, Charlotte S.; Montoya, Encarni; Valencia, Bryan G.; Williams, Joseph J.

Published in:
Palaeogeography, Palaeoclimatology, Palaeoecology

DOI:
10.1016/j.palaeo.2014.05.001

Publication date:
2014

Citation for published version (APA):
Keen, H. F., Gosling, W. D., Hanke, F., Miller, C. S., Montoya, E., Valencia, B. G., & Williams, J. J. (2014). A statistical sub-sampling tool for extracting vegetation community and diversity information from pollen assemblage data. *Palaeogeography, Palaeoclimatology, Palaeoecology*, *408*, 48-59. https://doi.org/10.1016/j.palaeo.2014.05.001

Palaeogeography, Palaeoclimatology, Palaeoecology 408 (2014) 48–59

Contents lists available at ScienceDirect

Palaeogeography, Palaeoclimatology, Palaeoecology

journal homepage: www.elsevier.com/locate/palaeo

A statistical sub-sampling tool for extracting vegetation community and diversity information from pollen assemblage data

Hayley F. Keen [a,*], William D. Gosling [a], Felix Hanke [b], Charlotte S. Miller [a], Encarni Montoya [a], Bryan G. Valencia [a,c], Joseph J. Williams [a,d]

[a] Department of Environment, Earth and Ecosystems, Centre for Earth, Planetary, Space & Astronomical Research (CEPSAR), The Open University, Walton Hall, Milton Keynes MK7 6AA, UK
[b] Accelrys, 334 Cambridge Science Park, Cambridge CB4 0WN, UK
[c] Department of Biology, Florida Institute of Technology, 150 West University Boulevard, Melbourne, FL 32901, USA
[d] Department of Geography and Earth Sciences, Aberystwyth University, Aberystwyth, Ceredigion SY23 2DB, UK

ARTICLE INFO

Article history:
Received 23 December 2013
Received in revised form 28 April 2014
Accepted 2 May 2014
Available online 9 May 2014

Keywords:
Palynology
Count size
Evenness
Richness
Sub-sampling
Tropical

ABSTRACT

Pollen assemblages are used extensively across the globe, providing information on various characteristics of the vegetation communities that originally produced them, and how these vary temporally and spatially. However, anticipating a statistically based robust pollen count size, sufficient to characterise a pollen assemblage is difficult; particularly with regard to highly diverse pollen assemblages. To facilitate extraction of ecologically meaningful information from pollen assemblage data, a two part statistical sub-sampling tool has been developed (Models 1 and 2), which determines the pollen count size required to capture major vegetation communities of varying palynological richness and evenness, and the count size required to find the next not yet seen (rare) pollen taxa. The sub-sampling tool presented here facilitates the rapid assessment of individual pollen samples (initial information input of 100 pollen grains) and can, therefore, on a sample by sample basis achieve maximum effectiveness and efficiency. The sub-sampling tool is tested on fossil pollen data from five tropical sites.

Results demonstrate that Model 1 predicts count sizes relating to palynological richness and evenness consistently. To characterise major vegetation community components model 1 indicates that, for samples with a lower richness and higher evenness lower count sizes than are considered standard can be used (<300, e.g. 122); however, for samples of high richness and low evenness, higher count sizes are required (>300, e.g. 870). Model 2 calculates the additional number of pollen grains needed to be counted to detect the next not yet seen pollen taxa, outputs were strongly related to input data count size as well as richness and evenness characteristics. We conclude that, given the temporal and spatial variations in vegetation communities and also pollen assemblages, pollen count sizes should be determined for each individual sample to ensure that effective and efficient data are generated and that detection of rare taxa is checked iteratively throughout the counting process.

1. Introduction

Fossil pollen contained within natural sedimentary records can be used to reconstruct past vegetation communities and assess how they have changed through time. The type of ecological information extracted from fossil pollen records includes: i) identifying large-scale shifts between biomes (defined here as a large array of flora and fauna within one major habitat), e.g. shifts between woodland and grassland (Rull et al., 2005), or shifts between deciduous forest and boreal forest (Fréchette and de Vernal, 2013), ii) determining first arrival or introduction of species (Hooghiemstra and Cleef, 1995; Van der Knaap et al., 2012), and iii) characterising shifts in criteria important for conservation, e.g. assemblage richness or the discovery of rare taxa (Bush and Colinvaux, 1988). Furthermore, examination of modern pollen–vegetation relationships can be used to address biogeographic and ecological questions (Jantz et al., 2014). Understanding the nature and dynamics of vegetation communities through time and space is essential in order to anticipate the likely response of modern vegetation to human activity, and on-going/projected climate changes (Jackson, 2012).

To be confident that the inferences being drawn about vegetation communities from the pollen assemblage data are valid, it is necessary to consider several key factors including: i) pollen production (Bush, 1995; Gosling et al., 2005), ii) pollen transport (Gosling et al., 2009; van der Knaap, 2009), iii) pollen preservation (Havinga, 1964, 1984), iv) distribution of pollen grains on the slide (Brooks and Thomas, 1967; Holt et al., 2011), v) taxonomic classification of pollen types and the relationship with the taxonomic classification of the parent vegetation (Odgaard, 1999), and vi) efficiency of sampling (Rull, 1987; Moore et al., 1991). Furthermore, consideration of other proxies (e.g.

* Corresponding author. Tel.: +44 1908 655298.
E-mail address: hayley.keen@open.ac.uk (H.F. Keen).

http://dx.doi.org/10.1016/j.palaeo.2014.05.001

macrofossils) is also important, as these can provide further insight into the vegetation communities (Birks and Birks, 2006). In this paper we focus on sampling efficiency and present a new methodology for sub-sampling (by means of pollen counting) pollen assemblages. The method presented allows the researcher to: i) tailor their sampling strategy to the scientific question being asked, and ii) account for variation in the pollen assemblages throughout time and space. This method ensures that if a vegetation change occurs, a statistically robust count size appropriate for its detection will be achieved.

2. Considerations for pollen counting

To establish a robust link between fossil pollen data and past vegetation communities, it is necessary for the researcher to consider the question(s) posed and balance the investigator effort required (time consumed), against time available. Next we consider two key factors related to effective and efficient pollen counting: i) determining an appropriate pollen count size, and ii) the application of the determined pollen count size to a study site.

2.1. Determining pollen count size

Research based on percentage rarefaction curves of pollen assemblages from temperate regions indicates that pollen count sizes (target amount of pollen grains to count within a single sample) between 300 and 500 grains (excluding aquatic taxa) are often enough, dependent on the question being investigated (Birks and Birks, 1980). However, larger count sizes (>500 grains) have also been recommended as a more suitable count size to characterise past vegetation composition (Moore et al., 1991; Bennett and Willis, 2001). Within more floristically diverse tropical regions, studies establishing an effective pollen count size are scarce (Rull, 1987); although pollen counts of >500 grains were found to be sufficient to characterise the major components of pollen assemblages in a study of modern pollen–vegetation relationships in Neotropical forests and savannahs (Gosling, 2004; Gosling et al., 2005).

Most investigations into past vegetation change are concerned with large scale characterisation of the vegetation and, therefore, pollen sums of >300 grains are widely used (following Birks and Birks, 1980); i.e. 74% percent of the top 50 most cited papers returned within Scopus (http://www.scopus.com) for the search term 'Quaternary fossil pollen' indicated that count sizes of at least 300 were targeted (25th October 2013). However, variance in either the richness (amount of taxa within an ecosystem), or evenness (representation of taxa within an ecosystem) of how the parent vegetation is expressed in the pollen assemblage could result in pollen sums of >300 being insufficient or in excess. The potential for variance in richness and evenness to hinder pollen counting accuracy is of particular concern when trying to reconstruct past vegetation from the tropics, mainly due to the high floristic diversity within these ecosystems. Consequently, tropical vegetation is more difficult to reconstruct (Odgaard, 2001). For example, in a fossil pollen sub-sample from the tropical eastern Andean flank, diversity characteristics (Fig. 1A) and relative taxon abundances (Fig. 1B) are shown to vary markedly dependent on count size.

2.2. Applying pollen count sizes to a study site

Currently, standard research practice is to apply an identical count size target for an entire study i.e. throughout a sedimentary sequence for fossil pollen, or across a series of vegetation plots for modern pollen studies. In a setting where the richness and/or evenness characteristics of the various vegetation communities being examined are roughly similar, the application of a uniform count size is more applicable, e.g. comparison of two types of temperate forest. However, if the time period, or area, being studied covers a shift in richness or evenness characteristics of the parent vegetation community, the use of a single count size could result in false inferences due to under or over sub-sampling (hereafter interchangeable sample/sub-sampling), and although over sampling is not a statistical issue, it does mean wasted investigator effort. For example, to characterise the major components of a savannah with low palynological richness, a lower pollen count size would be required in comparison to a palynologically diverse tropical forest (high richness). Richness changes could occur within an individual sedimentary sequence or study region, making it important to identify a count size appropriate for each sample. Consequently, a methodology is required to determine appropriate pollen count sizes on a sample-by-sample basis.

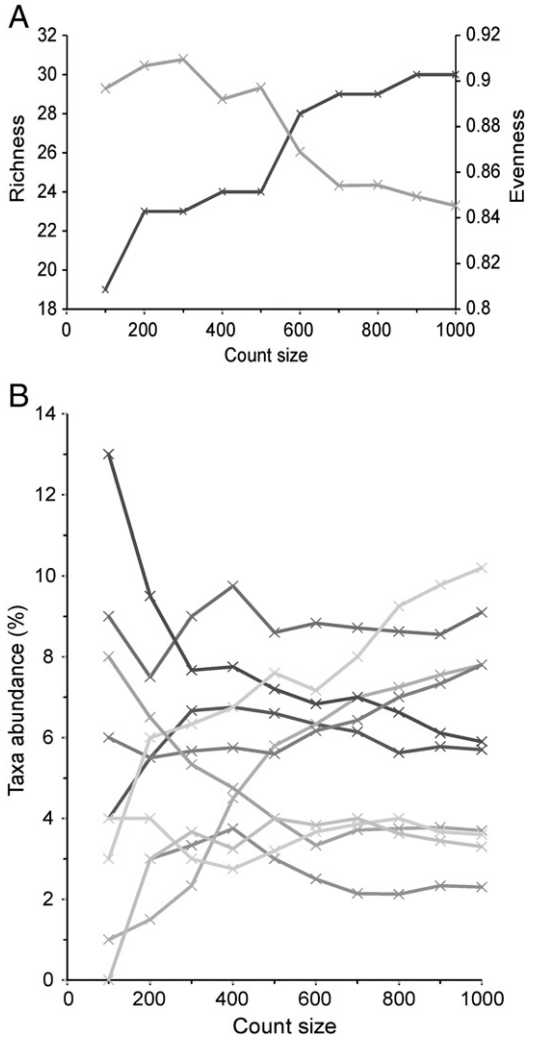

Fig. 1. Ecological descriptors of a fossil pollen assemblage from the tropical eastern Andean flank (Mera Tigre East) as a function of increasing sample size (increments of 100 grains). A) Diversity: total sub-sample richness (black line) and evenness (grey line). B) Assemblage composition: percentage cumulative taxa abundance for ten selected taxa.

2.3. Improving statistical sub-sampling of pollen assemblages

In this paper we present a statistical methodology (sub-sampling tool), which allows preliminary pollen count data to be used to assess the ideal pollen count size required to address three ecological questions: i) what are the major components of the parent vegetation community (biome), ii) what is the richness of the sample (including rare taxa), and iii) when is it probable that the next not yet seen pollen grain has been sampled? The statistical model presented takes into account the richness and evenness of a sample through the input of an initial pollen count of 100 pollen grains (see Section 5.1). To test the robustness of the model, extended fossil pollen count data from three tropical regions are compared against the model output: i) high

elevation central Andes (Bolivia and Peru), ii) mid-elevation eastern Andean flank (Ecuador), and iii) lowland West Africa (Ghana).

3. Ecological parameters

3.1. Richness

For the purpose of this paper, richness (R) is defined as the amount of taxa within an ecosystem and it is calculated as the total number of palynological taxa within a sample. R can only be used as a measure of palynological richness if the pollen count numbers are standardised, as such a fixed amount of grains (100 grains — see Section 5.1) is used to enter data into the statistical model so that R can be calculated. Rarefaction analysis can be used for standardisation purposes if the count sizes are different (Birks and Line, 1992).

In general for any given pollen assemblage a parent vegetation community with a high richness will have a greater number of different pollen taxa than a vegetation community with a low richness. However, due to the variances in pollen production for each taxon, the pollen assemblage does not always directly reflect the richness of the parent vegetation (Bush, 1995; Odgaard, 2001). Some taxa present in the vegetation can be underrepresented in the pollen assemblage, or even be absent (e.g. Orchidaceae), and consequently, the vegetation community richness may not be fully represented by pollen richness (Odgaard, 1999; Goring et al., 2013). Conversely long-distance transport of pollen grains into the study site from extra-regional vegetation communities could result in an artificially elevated palynological richness in comparison to the local parent vegetation community (Gosling et al., 2009).

3.2. Evenness

For the purpose of this paper, evenness (E) describes the distribution of pollen taxa within the pollen assemblage. In this sense, a sample with a dominant taxon (high number of pollen grains of the same type) would be considered low evenness, whereas a sample without dominant taxa (pollen grains similarly/equally distributed amongst all taxa) would represent high evenness (Smith and Wilson, 1996). In a pollen assemblage, understanding evenness is not always straightforward as different taxa produce varying amounts of grains and distribute pollen grains differently, e.g. anemophilous vs. entomophilous taxa (Sugita, 1994; Bush, 1995; Gosling et al., 2009). The variance in production and distribution can occasionally lead to difficulties in linking evenness within a pollen assemblage, to that of the parent vegetation. Nevertheless, the sub-sampling tool uses evenness in the pollen assemblage, so this does not affect model performance, i.e. the relationship between palynological evenness and the parent vegetation evenness still needs to be considered as usual when interpreting the pollen and vegetation relationship.

Evenness (E) is calculated using the following formulae (Eqs. (1)–(3)), where two variables are required: i) richness (R) (defined in Section 3.1) which is not heavily dependent on evenness, and ii) the Shannon–Wiener index (H), which is an index used to measure biodiversity, and is strongly influenced by evenness within the pollen assemblage.

$$P_i = \frac{\text{number of grains for an individual taxa}}{\text{total number of grains for all taxa}} \quad (1)$$

P_i is the proportion of a total sample belonging to the ith taxa (Krebs, 1999), and it is a variable used within the calculation of H.

$$H = -\sum_{i=1}^{S} (P_i * \ln[P_i]) \quad (2)$$

where Σ represents the sum, S is the number of taxa within the sample, i is a single taxa within the sample, ln represents the natural logarithm, and P_i is as defined above (Eq. (1)). Eq. (2) must be applied to all taxa individually and the summation of these calculations used for the calculation of H for the entire pollen assemblage.

Once R and H have been calculated the following formula can be applied to produce an E (evenness) value.

$$E = \frac{H}{\ln(R)} \quad (3)$$

The value of E can vary between 0 (low evenness) and 1 (high evenness).

4. Methodology

In order to develop and verify a robust statistical methodology for determining appropriate count sizes for pollen assemblages of varying richness and evenness, the following steps were applied: i) selection of study sites with vegetation of varying richness and evenness, ii) generation of empirical data (pollen preparation, identification and counting), iii) generation of statistically modelled pollen counts (subsampling tool), and iv) consideration of how to apply count size estimations to address particular ecological questions.

4.1. Study sites

To capture a wide range of evenness and richness values within pollen assemblage data, ten samples were analysed from three different tropical regions: i) high central Andes, ii) eastern Andean flank, and iii) lowland West Africa. Sites from high (three sites), mid (one site) and low (one site) elevations were selected for study to provide insight into a range of tropical vegetation communities. One sample was analysed for each site, except for the mid elevation site where six samples were analysed to investigate variance within one region through time. All samples were obtained from sedimentary sequences (fossil pollen records).

4.1.1. High elevation, central Andes, Bolivia and Peru, South America

Three high elevation study sites (Lakes Khomer Kotcha Upper, Challacaba and Pacucha) were used in this study and provide an opportunity to test the model output against a range of different richness and evenness values. The sediment cores from all three high elevation sites were collected using a Colinvaux modified Livingstone corer (Valencia et al., 2010; Williams et al., 2011a,b).

Khomer Kotcha Upper is a glacier formed lake situated in Bolivia (17°16.514′S, 65°43.945′W, 4153 m asl [above sea level]). Today the site has a mean annual temperature (MAT) of 4.5 °C to 7.6 °C and mean annual precipitation (MAP) of 772 mm. Modern vegetation present at the region transitions between puna grassland and punean woodland (Williams et al., 2011a). The sample chosen for this study from Khomer Kotcha Upper is from the Early Holocene (c. 9360 cal yr BP [calibrated years before present]).

Challacaba is a freshwater lake located in the Andes of Bolivia (17°33.257′S, 65°34.024′W, 3400 m asl). The MAT of the site varies from 7.2 °C to 11.3 °C annually and the precipitation varies seasonally between 2.6 mm and 114 mm per month. Current vegetation at the site is a patchwork of grassland, shrub and *Polylepis* sp. dominated woodland (Williams et al., 2011b). The sample chosen for this study from Challacaba is from the Late Holocene (c. 3270 cal yr BP).

Lake Pacucha is located in the Peruvian Andes (13°36.384′S, 73°19.690′W, 3095 m asl). MAT is 13 °C and MAP is <700 mm (Valencia et al., 2010). Human activity around the lake has resulted in a shift from native *Polylepis* sp. woodland to *Eucalyptus* sp. plantations and crops (mainly potatoes and barley). The sample chosen for this study from Pacucha is from the Last Glacial Maximum (c. 22,400 cal yr BP).

4.1.2. Mid elevation, eastern Andean flank, Ecuador, South America

Mera Tigre East is located on the eastern Andean flank in the Pastaza province of Ecuador (01°27.546′S, 78°06.199′W, 1117 m asl). Today, the Mera region has a MAT of 20.8 °C and MAP of >4800 mm (Ferdon, 1950; Liu and Colinvaux, 1985), and diverse vegetation including different degrees of human disturbed rainforests. Sediments were recovered from an 8.49 m vertical section exposed by the down cutting of the Rio Tigre. The six samples selected for this study are all of Pleistocene age (younger than 1 Ma due to the presence of *Alnus* in the pollen assemblage (Hooghiemstra, 1984), but beyond the limit of radiocarbon dating, i.e. >50,000 years old). The six samples from Mera Tigre East were selected for analysis because they presented an opportunity to test the model against multiple samples with high palynological diversity (richness).

4.1.3. Low elevation, central Ghana, West Africa

Lake Bosumtwi is located in the lowlands of Ghana, Africa (6°30′ N, 1°25′W, 97 m asl). The MAT is 26 °C and the MAP is 1260 mm (Shanahan et al., 2008). Prior to the degradation of the natural vegetation by human settlement and cultivation, the lake was surrounded by moist semi-deciduous forest (Gill, 1969), with a dominant canopy comprised of trees from the Ulmaceae and Sterculiaceae families (Hall and Swaine, 1981; Beuning et al., 2003). In 2004, 1833 m of sediments were recovered from Bosumtwi as part of the International Continental Drilling Program (Koeberl et al., 2007). The sample chosen from Bosumtwi was from the last glacial period (Miller and Gosling, 2014). The glacial sample from Bosumtwi was selected for analysis because of its low palynological richness.

4.2. Pollen preparation and identification

Pollen preparation at all sites followed standard procedure, including acetolysis and digestions with Hydrochloric acid, Potassium hydroxide and Hydrofluoric acid (Moore et al., 1991). Samples were spiked with an exotic marker to: i) allow the calculation of pollen concentrations (Stockmarr, 1971; Maher, 1972), and ii) provide a reference marker for the extended pollen counts. The samples were mounted on slides using glycerol and pollen was identified from their distinguishing morphometric features using reference material held at The Open University and Florida Institute of Technology, open access online pollen databases (Bush and Weng, 2007; Gosling et al., 2013), and published pollen atlases (Hooghiemstra, 1984; Roubik and Moreno, 1991; Reille, 1995; Colinvaux et al., 1999; Vincens et al., 2007; Gosling et al., 2009).

4.2.1. Extended pollen counts

Extended pollen counts for the six Mera Tigre East samples followed a protocol designed to assist in the development of the statistical model. Samples were counted until a total of 300 exotic markers (*Lycopodium* sp. spores, batch 124961, Lund University) were reached. *Lycopodium* spores were counted in batches of twenty and all terrestrial pollen grains were counted within the batches. Once a count of 300 *Lycopodium* spores had been achieved the analysis stopped and the percentage abundance for the terrestrial pollen grains was calculated. By counting to 300 *Lycopodium* marker spores, roughly 2000 terrestrial pollen grains were counted for each sample.

Lakes Khomer Kotcha Upper, Challacaba and Bosumtwi were extensively counted until a total of 1000 terrestrial pollen grains had been reached as suggested by Moore et al. (1991). Lake Pacucha was not counted extensively, but a 'standard' count of 300 pollen grains was achieved.

4.3. Modelled pollen counts

The statistical model (sub-sampling tool) requires the input of empirical pollen count data for each pollen sub-sample being considered. The statistical model generates multiple simulations of the possible permutations for pollen assemblages based on the input data, and then assesses the most probable count size required to capture the ecological characteristics (assemblage composition and diversity).

The primary aim of the model is to estimate the required count size that would reliably characterise the major components of the parent vegetation community (hereafter Model 1). The secondary aim of the model is to estimate how many more pollen grains would be needed before a not yet found pollen type was detected, i.e. a taxa not already seen is found (hereafter Model 2). The count size estimates will depend on the richness, as well as the evenness, of the sample based on the initial input data. However, it is important to note that the outputs are modelled probabilities, and that there are many complexities that mean the translation of modelled estimates to 'real world' pollen assemblages will not be perfect. The purpose of the model is, therefore, to provide statistically-based support for the researcher to help ensure the data generated have the best possible chance of addressing the question(s) posed.

4.3.1. Model methodology

The statistical model (sub-sampling tool) takes empirical pollen assemblage input data and runs a simple Monte Carlo simulation (a method involving running a simulation multiple times to assess the likely outcomes) one hundred times in succession. The multiple model runs allow the pollen count required to determine the ecological characteristics (assemblage composition and diversity) within a sub-sample to be estimated. The target amount of empirical pollen count data that should be examined to capture this can then be calculated.

The model pollen count simulations work by assigning each pollen grain a random number. For a given distribution of taxa the simulation then has to determine which taxa corresponds to the chosen random number. By repeatedly choosing random numbers, the model mimics the empirical pollen counting process.

In practice, the pollen count model procedure is as follows. At the beginning of each run, the relative abundance of each taxon is obtained (p_i), ensuring that the sum of p_i is 1. A cumulative amount $r_i = p_i + r_{i-1}$ (sum of all p_j for $i < j$) is determined for each taxon. Each modelled pollen grain is described by obtaining a single random s number between 0 and 1. The taxa index i of this pollen grain is obtained by working out which cumulative fraction corresponds to the random number, e.g. by testing for $p_i < s < p_{i-1}$. By repeating this process for each pollen grain, a randomised counting is achieved, corresponding to the suggested input data when a sufficient amount of pollen grains are counted.

Once these data have been generated within the model environment, the model then 'bins' the simulation data. Binning is a process simply used to group individual data values into one place; this is instead of displaying a large amount of data separately. In this statistical model, the simulation data can be binned using two different methods, either by a pre-defined number of pollen grains per bin, or by counting up to a fixed number of exotic markers in each bin. The method of binning can be chosen when using the model. The operator chooses a binning method and then 'turns off' the other method, meaning only one is used during any one model run. For the model runs used to generate the results in Figs. 2–6, the simulation data were binned using a fixed number of exotic markers (*Lycopodium* sp. spores) per bin. The other method, using a predefined number of pollen grains per bin, is useful if a low number of exotic markers have been counted, i.e. when the pollen concentration is high.

A single model count run takes no more than a few seconds, therefore, generating multiple model runs simultaneously is possible, and so statistical information can be obtained from the aggregated output. This allows the multiple runs necessary for the Monte Carlo simulation to be produced in a minimal time (c. 20 s). Data from multiple model runs allow an error estimate for the counted distribution to be calculated and, therefore, a pollen count size suitable for each specific sample to within a 95% confidence interval to be determined (see Section 8 for further information on output data).

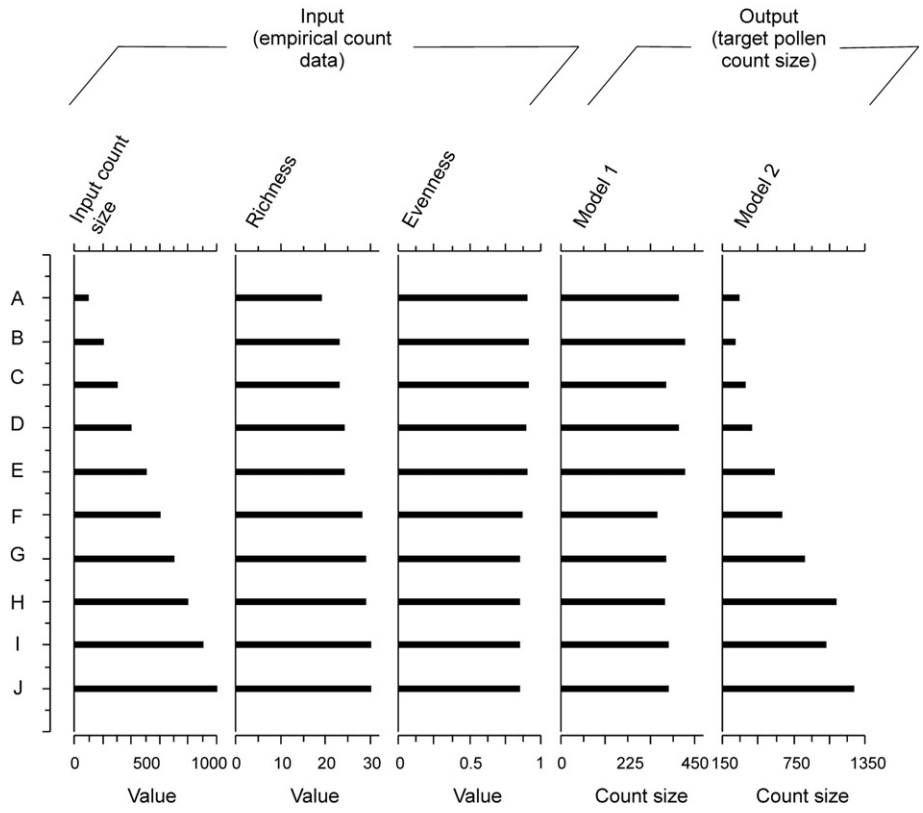

Fig. 2. Model response to increasing amount of pollen count input data from a fossil pollen assemblage from the eastern Andean flank (Mera Tigre East). The amount of pollen count input data varies from 100 grains (sub-sample A) to 1000 grains (sub-sample J) in increments of 100 grains. Count size outputs for detecting major vegetation composition (biome) Model 1 (Section 4.4.1) and the next not yet seen pollen taxa Model 2 (Section 4.4.2) for each sample are shown.

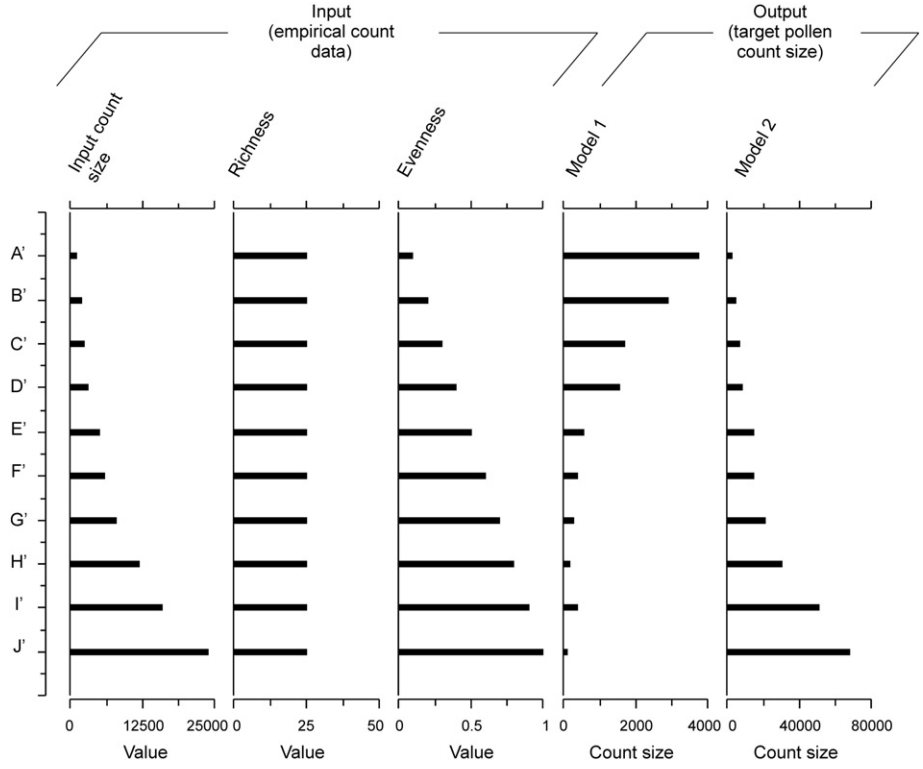

Fig. 3. Model response to increasing evenness in pollen assemblage composition. The richness value was kept the same (25) and the evenness increased from 0.1 (sub-sample A′) to 1.0 (sub-sample J′) in increments of 0.1. To maintain richness and evenness values it was necessary to use different input count size values. Count size outputs for detecting major vegetation composition (biome) Model 1 (Section 4.4.1) and the next not yet seen pollen taxa Model 2 (Section 4.4.2) for each sample are shown. Data used is from a random generation and is not indicative of any of the study sites.

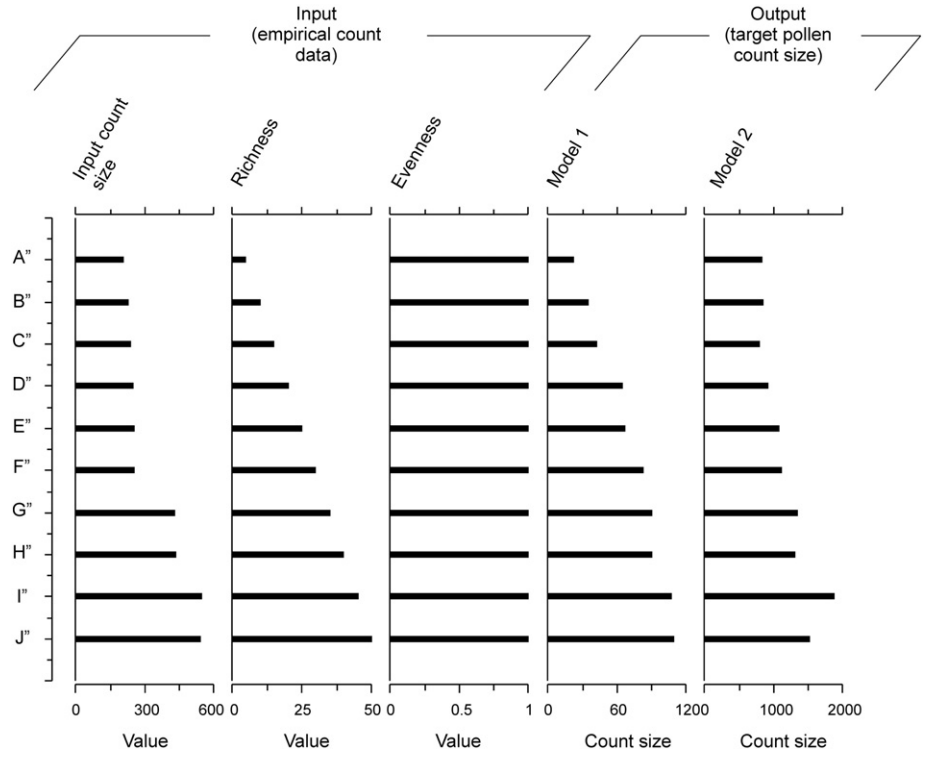

Fig. 4. Model response to increasing richness in pollen assemblage composition. The evenness value was kept the same (1.0) and the richness increased from 5 (sub-sample A″) to 50 (sub-sample J″) in increments of 5. To maintain richness and evenness values it was necessary to use different input count size values. Count size outputs for detecting major vegetation composition (biome) Model 1 (Section 4.4.1) and the next not yet seen pollen taxa Model 2 (Section 4.4.2) for each sample are shown. Data used is from a random generation and is not indicative of any of the study sites.

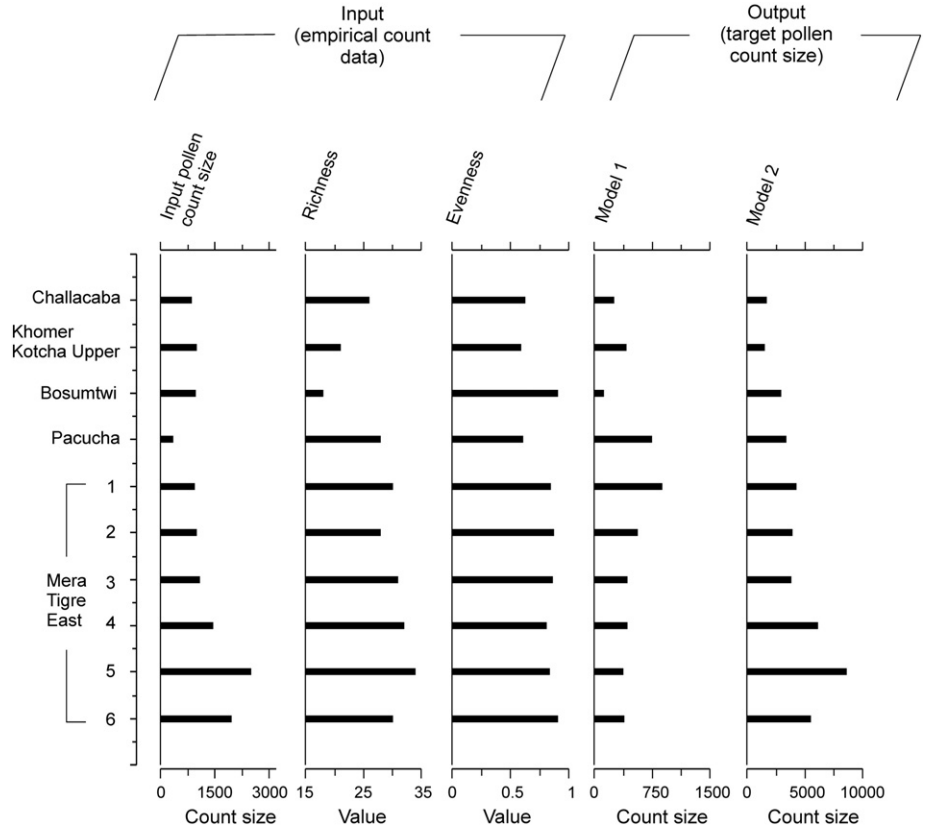

Fig. 5. Model count size estimates for ten fossil pollen assemblages obtained from five different tropical study sites. Each fossil pollen assemblage has different ecological characteristics (richness and evenness). The pollen count size outputs for detecting major vegetation composition (biome) Model 1 (Section 4.4.1) and the next not yet seen pollen taxa Model 2 (Section 4.4.2) for each sample are shown, alongside the empirical pollen count achieved through extended pollen counting.

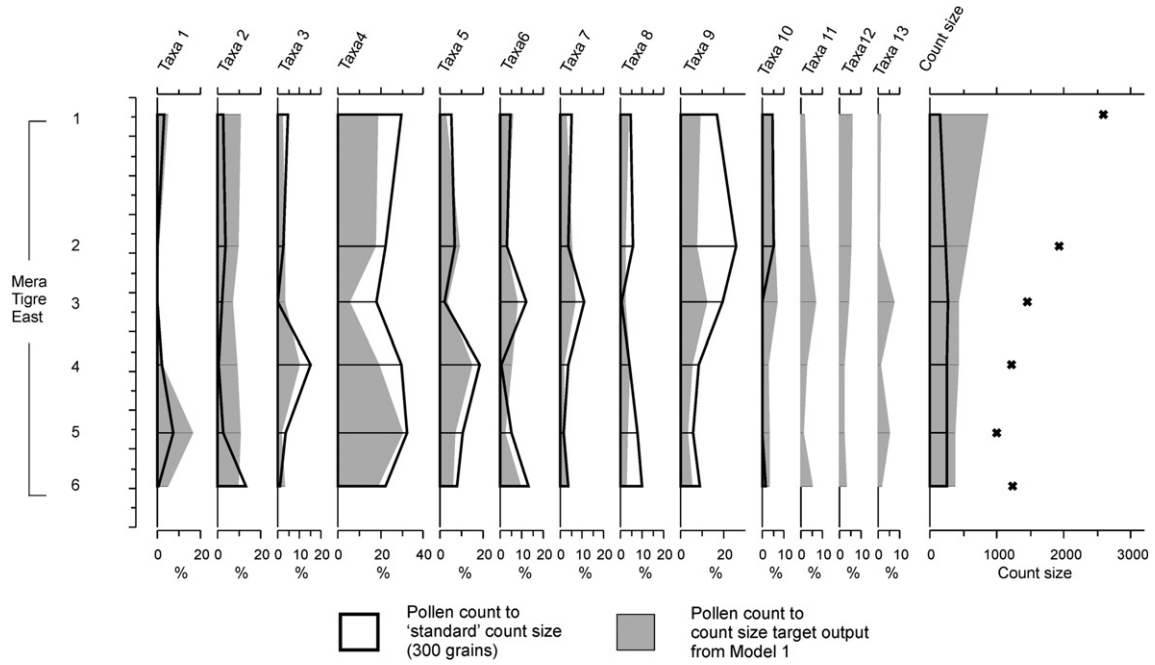

Fig. 6. Pollen assemblage data from a mid-elevation site on the eastern Andean flank (Mera Tigre East) for pollen count sizes of: i) 300 grains (black outline) and ii) extended counts based on statistical sub-sampling tool estimates (Model 1), count size estimates required to detect the next not yet seen taxa (Model 2) are also shown (black crosses); Model 2 estimates are based on an input of the extended pollen assemblage count data generated when counting to the target output from Model 1. All taxa >5% abundance are shown.

4.3.2. Model parameters

For each model run the sub-sampling tool simulates the equivalent of an empirical pollen count size of 2000 grains. The model is then run one hundred times in succession (multiple simulations needed for the Monte Carlo simulation) for the individual sub-sample being considered; the equivalent of considering the possible combination of 200,000 pollen grains. Generating an empirical pollen count of 200,000 grains from any one sample would take weeks, and so, is impractical. Therefore, the model provides an opportunity to explore the characteristics of pollen assemblage data which was not previously practical.

4.4. Determining appropriate count sizes for specific scientific questions

Two different statistical approaches are presented to address three questions (listed in Section 2.3) which can be asked of the pollen data. The first statistical approach (Model 1) determines the probability that the pollen count has correctly characterised the major components of the pollen assemblage (Question i). The second (Model 2) assesses the likely investigator effort required (number of additional pollen grains that must be counted) to detect the next not yet seen taxa within the pollen assemblage (Questions ii and iii).

4.4.1. Characterising major vegetation components (Model 1)

To determine statistically if the major components of the pollen assemblage have been characterised, the rank abundance of taxa within a sample was examined, i.e. has the pollen count been of a sufficient size to arrange the major components of the pollen assemblage in the 'correct' order. To assess the rank abundance, the Spearman's rank correlation coefficient (a method of assessing the link between two different variables) was calculated for a series of modelled count sizes (100, 200, 300 grains continuing up to 1000 grains) and compared against the model endpoint (equivalent count size 200,000 pollen grains).

Once a Spearman's rank value of over 0.95 (standard statistical 95% confidence level) is attained, then a reliable count is considered to have been reached. Therefore, when the determined count size has been achieved it can be considered that all major taxa have been correctly characterised in terms of rank abundance, i.e. characterised when the relative abundance of each of the major taxa (>5% abundance) has met the proposed model proportion estimate within the sample. This way of establishing a vegetation community representation in the pollen record (through means of major taxa) is used hereafter. However, in certain circumstances (e.g. very high richness) the Spearman's rank correlation coefficient can never reach 0.95. This is because although major components of the rank correlation will look identical, small differences in low abundance (minor) taxa can mean the Spearman's rank correlation coefficient will not increase to a value of 0.95. In this scenario the standard deviation of the Spearman's rank correlation coefficient can be considered, i.e. the major components have been characterised, but there is still uncertainty within the minor taxa, leading to lower Spearman's rank correlation coefficient values. To circumvent the statistical problem caused by the low abundance taxa, a secondary threshold has been established at the point when the standard deviation of the Spearman's rank correlation coefficient reaches less than 0.05 for all taxa. The standard deviation determines how far the Spearman's rank correlation coefficient value deviates from the average. Once the standard deviation has reached a low level (<0.05), it indicates that the relative proportions of the majority of the taxa within the pollen assemblage have been successfully determined. The standard deviation is calculated automatically within the model and it is taken across all of the model runs and calculated for each Spearman's rank coefficient, for each of the separate count size bins. Unfortunately, it is not possible to calculate standard deviations for empirical pollen count data in the same way because it is not practical to generate an equivalent quantity of data.

4.4.2. Determining sample richness and detecting rare and first occurrence of taxa (Model 2)

To address ecological questions regarding diversity, and for the detection of rare taxa, a different form of statistical assessment is required, i.e. to determine how many more pollen grains would have to be examined to discover the next as yet unseen pollen taxa. An estimation of the investigator effort required to detect the next 'missing' taxa from any sub-sample, can be obtained as follows.

The unknown taxa is defined as taxa *x*, something likely to be there, but not yet discovered. If the aim is to find taxa *x* within the sample, then a high enough total count (N_{tot}) needs to be achieved. This experimental set up follows a Poisson distribution (probability of a given number of the unknown taxa *x* occurring in a fixed interval, in this case a total count size), meaning the variance on the number of unknown pollen grains within a sample is xN_{tot} and the standard deviation of the unknown count size is $\sqrt{xN_{tot}}$. It is also known that within each pollen sample, there will be a number of difficult to identify pollen grains (N_{unc}), this could be known as the unknown taxa *x*. In many cases the value N_{unc} can be set to equal 1, simply corresponding to the next (unknown) pollen taxa at the end of the sample. The aim is to know to within *S* standard deviations how many pollen grains need to be counted for it to be likely that the number of unknown pollen grains in the sample is smaller than N_{unc} given a total N_{tot} and an unknown *x*. These numbers have to satisfy the following inequality (Eq. (4)) which can then be solved as an equality for xN_{tot} (Eq. (5)).

$$N_{unc} < xN_{tot} - S\sqrt{xN_{tot}} \tag{4}$$

To solve this inequality, solving for $S\sqrt{xN_{tot}}$ followed by squaring both sides will give a quadratic equation in xN_{tot} with the following solution:

$$xN_{tot} > N_{unc} > + \frac{S^2}{2} + \frac{S}{2}\sqrt{S^2 + 4N_{unc}} \tag{5}$$

If it is hypothesised that we have no more than one pollen unidentified taxa (e.g. the next not yet seen taxa) then, to within one standard deviation *S*, the number of N_{tot} counts required to ensure that unknown taxa make up no more than a fraction *x* of the total sample is represented in Eq. (6).

$$xN_{tot} > \frac{3 + \sqrt{5}}{2} = 2.618 \tag{6}$$

As a simple example, let's consider a pollen sample in which, to one standard deviation, it is important to make sure that no more than 0.1% of pollen grains belong to an unknown taxa (i.e. $x = 0.001$ and $S = 1$). If every pollen grain can be identified, the required count size is $N_{tot} = 2.618/0.001 =$ a minimum count size of 2618 pollen grains. If there are N_{unc} unidentified grains, then the inequality (Eq. (5)) can be applied instead to find the required count size N_{tot} for a given fraction of unknown taxa *x*.

5. Results

Ideal count size estimates were produced from the examination of both empirical data generated from extended pollen counts and the sub-sampling tool outputs (Fig. 5).

5.1. Assessment of preliminary data input into the model

Pollen assemblage data for input into the model must be done at a consistent taxonomic level. Some plant families produce pollen grains which are hard to separate taxonomically (e.g. Poaceae), and thus are often classified only to family level within pollen assemblage data (Jantz et al., 2014). Keeping the taxonomic classification level of model input data standard between samples allows values obtained from different sub-samples to be compared. This does not mean that the palynologist needs to ignore the opportunity to classify pollen grains at a finer taxonomic level; it simply means that once classified and counted, the grains should be grouped into a consistent taxonomic level for input into the model for the different sub-samples being studied. If the grains in some sub-samples cannot be classified to the same taxonomic level, then taxa should be grouped together into the appropriate genus/family to keep the input consistent.

In order for each model run (each individual sample) to be comparable, it is also important to have comparable input variables, specifically, the amount of empirical pollen count data that are input into the model. A test run was performed using an extended count from the eastern Andean flank (Mera Tigre East) to see whether variations in the quantity of pollen count data input affected the model output target pollen count size. Due to the extended count being performed in stages, it was possible to input pollen grain counts of increasing increments (100, 200, 300 and so on.) into the model sequentially.

Regardless of how much empirical pollen assemblage data is input into the model, the model output count size does not fluctuate by more than ±13% from the average count size for Model 1 (Fig. 2). This indicates that Model 1 count size estimates from a preliminary input of 100 pollen grains are equally as acceptable as those generated when 1000 pollen grains are input. Therefore, to save time, and minimise wasted pollen counting effort, we recommend only 100 grains need to be initially counted for input if the detection of major vegetation components is the goal (Model 1). The target pollen count size estimate produced by the model can then be used to acquire the remaining empirical pollen assemblage data.

The amount of additional pollen data required to detect the next not yet seen pollen taxa, is directly related to the size of the empirical pollen count input (Fig. 2), i.e. when only 100 grains have been counted it is likely that you will come across a new taxa quicker than if you have counted 1000 grains (compare Fig. 2A with J). Therefore, for the detection of the next not yet seen pollen taxa Model 2 is required to be run iteratively through the counting process. At each step of the model, a decision can be made by the palynologist for the satisfactory completion of the sub-sample count. To ensure consistency between sub-samples, a threshold cut-off value could be set. For example, a sub-sample pollen assemblage count could be considered to have been completed when no further new taxa are anticipated to be discovered within the next 500 grains based on the estimates of Model 2.

5.2. Assessment of model effectiveness

To assess the functioning of the model, 'dummy' pollen assemblage data were input to ascertain how different ecological characteristics (e.g. richness and evenness) affected the model target pollen count size output estimates. These checks ensure that the model is working intuitively and give confidence for the application of the statistical sub-sampling tool as a guide for empirical pollen counting (see Section 8 for details of the operation of the model).

Two hypothetical pollen assemblage data sets of ten samples each were input into the model; the amount of pollen data input for each sample was varied to maintain the desired richness and evenness characteristics. In the first data set (scenario 1) the evenness increased whilst the richness remained consistent (Fig. 3). In the second dataset (scenario 2), the richness increased whilst the evenness remained consistent (Fig. 4). By using the two test scenarios it was possible to see, under controlled conditions, whether the model is performing as anticipated.

5.2.1. Model 1

In scenario 1 it was expected that as evenness increased, the order of the major pollen assemblage components should become easier to predict, i.e. the Model 1 estimate target pollen count size should decrease proportionally with increased evenness of the pollen assemblage data (Fig. 3).

In scenario 2 it was expected that as richness increases, the pollen assemblage should become increasingly harder to predict, i.e. the Model 1 estimate target pollen count size should increase proportionally with increased richness of the pollen assemblage data (Fig. 4).

The two hypothetical pollen count data sets input into the statistical sub-sampling tool demonstrate that the model is performing intuitively, i.e. more data (higher pollen counts) are recommended for pollen assemblages with low evenness and high richness characteristics.

5.2.2. Model 2

As anticipated from investigation of the input count size data (Fig. 2), in both scenarios 1 and 2 the outputs from Model 2 were closely related to the size of the pollen assemblage data input (Figs. 3 and 4).

5.3. Model output pollen count size estimates for specific study sites

To further assess model performance, model target pollen count size outputs were generated for ten samples from five different tropical fossil pollen records (Section 4.1). Modelled target pollen count sizes for major taxa (biome) characterisation (Model 1) were highest for Mera Tigre East 1 (870 grains) and Pacucha (741 grains), and lowest for Bosumtwi (122 grains) and Challacaba (256 grains) (Fig. 5). Modelled target pollen count sizes for rare taxa detection (Model 2) were highest for Mera Tigre East 5 (input data: count size = 2495, richness = 34, evenness = 0.83; output estimate of additional grains required to count to detect one new taxa = 8553) and Mera Tigre East 4 (input data: count size = 1449, richness = 32, evenness = 0.81; output estimate of additional grains required to count to detect one new taxa = 6116 grains), and lowest for Khomer Kotcha Upper (input data: count size = 1000, richness = 21, evenness = 0.59; output estimate of additional grains required to count to detect one new taxa = 1518 grains) and Challacaba (input data: count size = 860, richness = 26, evenness = 0.63; output estimate of additional grains required to count to detect one new taxa = 1667 grains) (Fig. 5). Different consideration of evenness and richness characteristics of the pollen assemblage within the two models means that the highest (lowest) pollen count size estimates from one model do not necessarily correspond to the highest (lowest) pollen count size estimates from the other (Fig. 5).

5.4. Example application of the statistical sub-sampling tool to a fossil pollen record

To assess if using the count size estimates produced by the model for major taxa detection (Model 1) made a difference to the reconstructed vegetation community from a fossil pollen record, two different pollen counts from each of the six Mera Tigre East samples were compared (Fig. 6). The first pollen count comprises pollen assemblage data obtained from a 'standard' pollen count of 300 terrestrial pollen grains (black outline). The second pollen count comprises pollen assemblage data obtained from pollen count targets estimated by Model 1 (ranging from 370 to 870 grains per sample; grey silhouette). The pollen assemblage data acquired from the two sub-sampling techniques indicate that 10 of the 13 the major taxa present (>5% abundance) are the same for both the 'standard' and 'model' count size target pollen assemblage data; however, the relative abundance of some taxa alters, e.g. decrease in relative abundance of taxa 4 and 9, and increase in relative abundance of taxa 1 and 2. The consistent occurrence of most of the major components demonstrates that the count size of 300 was just about sufficient to identify the major elements of the vegetation community, but not to establish the relative proportions.

Although Model 1 was not specifically designed to improve the detection of pollen assemblage richness, by counting to the higher target count sizes new insight into the diversity of the samples has been revealed. Three previously unidentified taxa are now recorded at >5%. Using the pollen assemblage data from the counts achieved following the guidance from Model 1 as input, the model was run again and an estimate of the additional grains to be counted to detect one more taxa was obtained from Model 2 (Fig. 6). Additional pollen counting of between 994 grains (sample 5) and 2598 grains (sample 1) were estimated by Model 2. Therefore, to detect the presence of any further taxa within the Mera Tigre East samples, significant additional investigator effort would have to be deployed.

6. Discussion

6.1. Application of statistical sub-sampling tool to study sites

To ensure that the maximum amount of information can be collected from a palynological investigation without counting an excess of pollen grains, it is recommended that each sample is treated individually. Individual pollen samples, regardless of whether or not they are from the same sedimentary sequence, or different study sites, are characterised by different richness and evenness values (Fig. 5). The models both take into account the richness and evenness characteristics of individual samples to produce an estimate of the optimal pollen count size required to: i) determine the relative abundance of major taxa (Model 1), and ii) detect the next not yet seen pollen taxa (Model 2).

Pollen assemblages with high richness require a pollen count size higher than the often used 'standard' pollen count target of 300 grains to characterise the major components. For example the high richness Mera Tigre East sample 1 (richness = 30) has an estimated target pollen count size of 870 grains (Model 1; Fig. 5). Application of the pollen count size recommended by Model 1 reveals a change in major taxa abundance and richness between counts of 300 and 870 grains (Fig. 6), i.e. a count size of only 300 grains for Mera Tigre East sample 1 provided an inaccurate picture of the major taxa abundance and an under representation of richness in the pollen assemblage.

Pollen assemblages with high evenness require a relatively lower pollen count size, because, with the increase in evenness the sample will be more easier to describe as all taxa are equally represented. The Bosumtwi study site has the highest evenness (and lowest richness) of all of the sites presented here (Fig. 5). The combination of high evenness and low richness results in the statistical sub-sampling tool estimating an ideal pollen count size of only 122 grains (Fig. 5). Therefore, if this Bosumtwi sample was counted to the 'standard' 300 terrestrial pollen grains then the palynologist would 'waste' effort on the sample.

As count sizes are predominantly driven by evenness and richness (and other factors mentioned in Sections 1 and 2), it is imperative that these are calculated (from the initial input data) for each sample and taken into account when using the statistical sub-sampling tool to generate target pollen count sizes. Application of the statistical sub-sampling tool for acquiring a target count size for pollen assemblage data will help ensure that investigator effort is efficient, without compromising the statistical robustness of the pollen assemblage data produced. The application of Model 2 to a data set to assess how efficiently pollen assemblage richness has been sampled is best done iteratively due to the relationship between probability of detection and pollen assemblage count size (Fig. 2).

The following steps are recommended for using the statistical sub-sampling tool as a guide for empirical pollen counting:

i) Count 100 pollen grains from sub-sample.
ii) Run model using count of 100 pollen grains as input data.
iii) Extract target pollen count size estimate from Model 1.
iv) Count sub-sample up to Model 1 pollen count size target.
v) Run model using data from Model 1 count size target as input data.
vi) Extract number of additional grains required to detect next not yet seen pollen taxa from Model 2.
vii) Evaluate if additional investigator effort is required/possible. If "no", pollen count of this sub-sample is complete. If "yes", iterate steps v–vii until the answer is "no" using increased pollen count sizes as input data for each increment.

6.2. Detecting major vegetation community (biome) composition assemblage data

As anticipated, Model 1 predicts count sizes which are higher for pollen assemblages with high richness and low evenness, e.g. for the Pacucha sample, which has a high richness (28 taxa in a pollen count of 337 grains) and low evenness (0.608) characteristics, a pollen count size of 741 grains is estimated (Fig. 5). In contrast, pollen assemblages which have lower richness and a higher evenness are estimated to require a relatively lower count size to detect major vegetation composition, e.g. for the Bosumtwi sample, which has low richness (18 taxa in a pollen count of 971 grains) and high evenness (0.9), a pollen count size of 122 grains is estimated. Estimates of pollen count sizes required to characterise the parent vegetation community in the Bosumtwi sample are much lower than the Pacucha sample, and much lower than the 'standard' 300 grain pollen count size that is widely used. Therefore, in the Bosumtwi example, the application of Model 1 to determine the pollen count size required during pollen counting would have reduced investigator effort into this sample, i.e. there was little point counting to >300 grains for the purpose of determining major vegetation components.

The six Mera Tigre East samples were analysed to both a 'standard' count size of 300 grains, and to the count size estimates produced by Model 1 (Fig. 6). After the Model 1 count size targets had been achieved, the most noticeable difference to the pollen assemblage was that the relative abundance of some major taxa changed (taxa 2, 4 and 9) and three new taxa were detected at >5% abundance (Fig. 6). Although most of the major taxa within the pollen assemblage were ever present, the change in relative abundances and diversity could impact the interpretation of the pollen assemblage data. In the Mera Tigre East example, counting to Model 1 estimates indicates a more diverse parent vegetation community

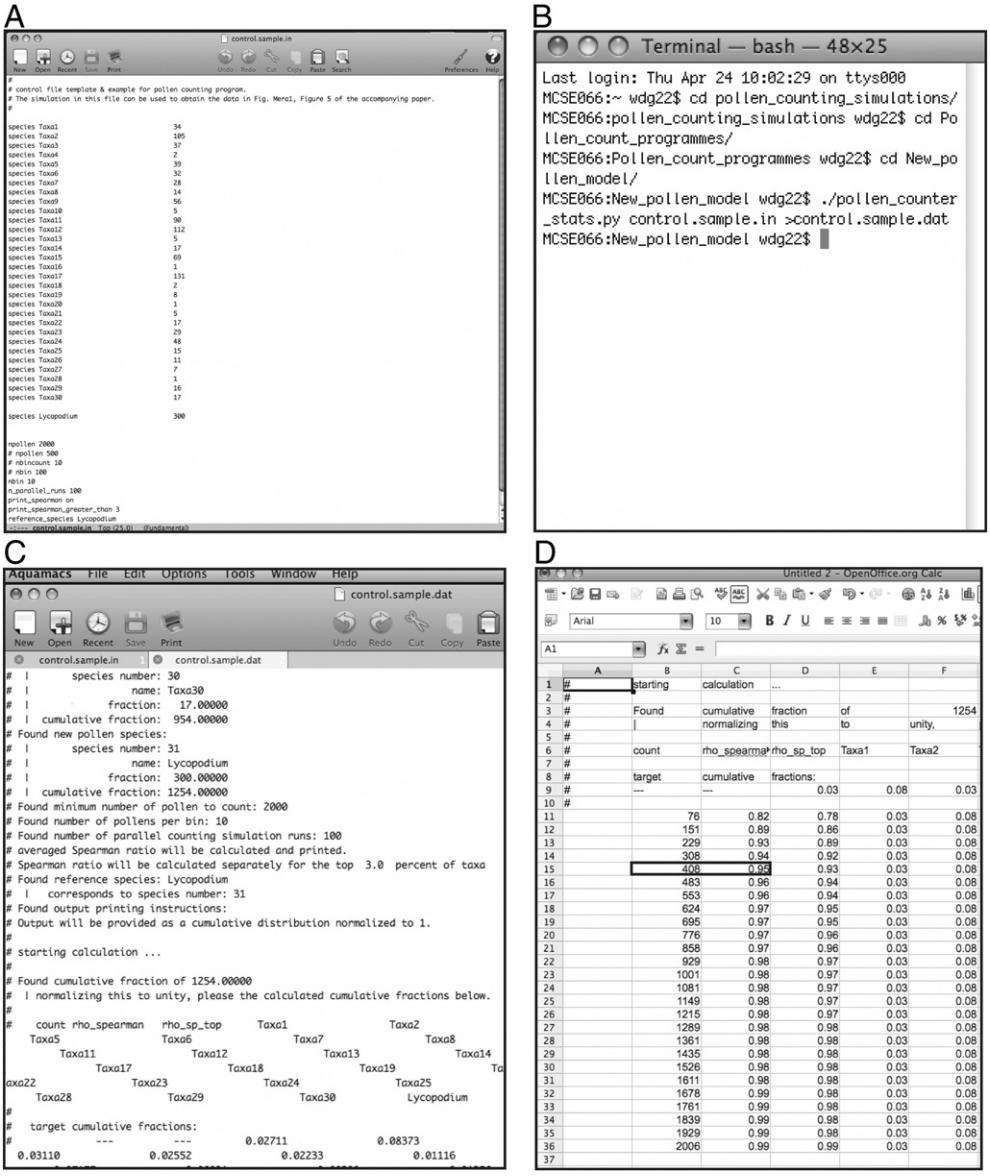

Fig. 7. Model input file for 'control.example.in' shown in Aquamacs (www.aquamacs.org). A) represents the format for data input into the model. The input file is composed in Aquamacs. Data should be formatted in three columns, column one being species, two being the taxa name and three being the count of the taxa. The marker should be included as a reference (here *Lycopodium*). B) represents the code required to run the model (./pollen_counter_stats.py control.sample.in >control.sample.dat), this is simply the model file, the input file name and then the name you want to output the file to. C shows the model output in the form of a.dat file (shown in Aquamacs). D) shows the model output exported into Microsoft Excel. The count size is reached when column C (rho_spearmans) reaches 0.95; in this case the count size is 408.

than would have been interpreted if only the 'standard' pollen counts of 300 grains had been achieved.

6.3. Detecting pollen assemblage richness and rare taxa

To provide statistical support for the detection, or otherwise, pollen count size estimates required to detect the next not yet seen pollen taxa are provided by Model 2. The probability of how quickly a 'new' pollen taxa will be detected by further counting is directly related to the number of pollen grains already counted and input into the model (Figs. 2–5). Consequently, to use Model 2 as a guide for pollen counting we advocate an iterative application (Section 6.1). Given the probability of extra-regional pollen being transported into any study site amongst other reasons, it is unlikely that any pollen count can ever sample the total pollen richness of an assemblage. Therefore, the decision on when a pollen count is complete becomes a trade-off between investigator time available and importance of detecting rare taxa to the ecological question posed.

The Mera Tigre East example was counted up to the target pollen counts established using Model 1 (Fig. 6, data in grey silhouette), and these data were then input into the model to obtain an estimate of how many more pollen grains would need to be counted for each sample to detect another pollen taxa (Model 2; Fig. 6, black and crosses). In all instances to reach the Model 2 target the pollen count size would have to be more than doubled to detect one further pollen type (Fig. 6). Therefore, in the Mera Tigre East example, it was decided that the high investigator effort required (more than doubling of time already invested), coupled with the low additional insight into the ecological characteristics projected (detection of one additional pollen type in each sample) was insufficient to merit further pollen counting. The Mera Tigre East fossil pollen samples were, therefore, considered to be complete upon reaching the pollen count size estimates projected by Model 1.

7. Conclusions

The widely used 'standard' pollen count size of 300 terrestrial pollen grains has been shown to be sufficient to provide an overview of pollen assemblage major composition (Fig. 1, (Birks and Birks, 1980). However, consideration of the ecological characteristics (richness and evenness) of individual pollen assemblages (sub-samples) can facilitate more effective and efficient pollen counting. The sub-sampling tool presented here offer an alternative methodology for pollen counting specifically designed to detect both vegetation community (biome) composition and richness.

We recommend that the statistical sub-sampling tool be applied to palynological investigations on a sample-by-sample basis to account for the variance in parent vegetation community (pollen assemblage) ecological characteristics through both time and space. We recommend that Model 1 be applied to palynological investigations interested in determining major components of vegetation communities. We recommend that Model 2 is only applied after target count sizes estimated by Model 1 have been achieved, and to palynological investigations where determining the diversity characteristics, and/or detection of rare taxa is particularly critical due to the high investigator effort required. The key advantages to the palynologist of using the statistical sub-sampling tool are:

- All pollen assemblage data have the same statistical confidence (not count size).
- Pollen count size targets are linked to the research question.
- Investigator effort can be deployed in a targeted manner.

Although designed with the specific application to investigate pollen assemblages, there is no reason why the statistical sub-sampling tool presented here could not be used to guide other types of ecological and palaeoecological investigations.

8. Statistical sub-sampling tool

README file for pollen counter python package.
This archive should contain six files:

- README: The current file.
- LICENSE: A copy of the LGPL v3.0 which governs the distribution of the program.
- pollen_counter.py: python executable for a single pollen counting run.
- pollen_counter_stats.py: python executable for a common pollen counting run with Spearman's rank statistics included.
- documentation.txt: Documentation of all available keywords for controlling the simulation.
- control.example.in: example simulation used for Fig. 6 in the accompanying paper. Fig. 7 represents the input and output files for this file. Note: This example contains more than the recommended 100 input grains.

NOTE: This package requires an implementation of the python programming language to be installed and needs to be run from a terminal prompt. Some operating systems (e.g. linux/unix and OSX) provide these facilities by default, others may require the download of additional software available freely on the internet.

Acknowledgements

This research was undertaken as part of a PhD studentship (HFK) funded by NERC (NE/J500288/1) and the Open University. Two anonymous reviewers are also acknowledged for their helpful comments which much improved the manuscript. The following authors contributed to the manuscript: HFK collated the data, counted the Mera Tigre East samples and wrote a large proportion of the manuscript. WDG wrote and edited a large proportion of the manuscript. FH developed the statistical model and contributed to the manuscript. CSM contributed the data for Bosumtwi and commented on improve the manuscript. EM helped with the writing of the manuscript. BGV contributed the data for Pacucha and commented on the manuscript. JJW contributed the data for Khomer Kotcha Upper and Challacaba and commented on the manuscript.

Appendix A. Supplementary material

Six files comprising the statistical sub-sampling tool will be uploaded as supplementary material. This will contain Model 1 and Model 2, one documentation file, the license, an example run file and an instruction document. It will also contain information about all of the software required to run the model. Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.palaeo.2014.05.001.

References

Bennett, K.D., Willis, K.J., 2001. In: Smol, J.P., Birks, J.H.B., Last, W.M. (Eds.), Pollen. Terrestrial, Algal and Siliceous Indicators, Volume 3. Kluwer Academic Publishers, Dordrecht.
Beuning, K.R.M., Talbot, M.R., Livingstone, D.A., Schmukler, G., 2003. Sensitivity of carbon isotopic proxies to paleoclimatic forcing: A case study from Lake Bosumtwi, Ghana, over the last 32,000 years. Glob. Biogeochem. Cycles 17, 31–32.
Birks, H.J.B., Birks, H.H., 1980. Quaternary Palaeoecology. , The Blackburn Press, New Jersey.
Birks, H.H., Birks, H.J.B., 2006. Multi-proxy studies in palaeolimnology. Veg. Hist. Archaeobot. 15, 235–251.
Birks, H.J.B., Line, J.M., 1992. The use of rarefaction analysis for estimating palynological richness from Quaternary pollen — analytical data. The Holocene 2, 1–10.
Brooks, D., Thomas, K.W., 1967. The distribution of pollen grains on microscopic slides. The non randomness of the distribution. Pollen Spores 9, 621–629.
Bush, M.B., 1995. Neotropical plant reproductive strategies and fossil pollen representation. Am. Soc. Nat. 145, 594–609.
Bush, M.B., Colinvaux, P.A., 1988. A 7000-year pollen record from the Amazon lowlands, Ecuador. Vegetatio 76, 141–154.

Bush, M.B., Weng, C., 2007. Introducing a new (freeware) tool for palynology. J. Biogeogr. 34, 377–380.
Colinvaux, P.A., De Oliveira, P.E., Patiño, J.E.M., 1999. Amazon Pollen Manual and Atlas, Harwood Academic Publishers, Amsterdam.
Ferdon, E.N.J., 1950. Studies in Ecuadorian Geography, School of American Research and University of Southern California, Santa Fe, New Mexico.
Fréchette, B., De Vernal, A., 2013. Evidence for large-amplitude biome and climate changes in Atlantic Canada during the last interglacial and mid-Wisconsinan periods. Quat. Res. 79, 242–255.
Gill, H.E., 1969. A ground-water reconnaissence of the Republic of Ghana, with a description of geohydrologic provinces. USGS water supply paper, 1757-K.
Goring, S., Lacourse, T., Pellatt, M.G., Mathewes, R.W., 2013. Pollen assemblage richness does not reflect regional plant species richness: a cautionary tale. J. Ecol. 101, 1137–1145.
Gosling, W.D., 2004. Characterisation of Amazonian forest and savannah ecosystems by their modern pollen spectra. (PhD), Department of Geography, University of Leicester.
Gosling, W.D., Mayle, F.E., Tate, N.J., Killeen, T., 2005. Modern pollen-rain characteristics of tall terra firme moist evergreen forest, southern Amazonia. Quat. Res. 64, 284–297.
Gosling, W.D., Mayle, F.E., Tate, N.J., Killeen, T.J., 2009. Differentiation between Neotropical rainforest, dry forest, and savannah ecosystems by their modern pollen spectra and implications for the fossil pollen record. Rev. Palaeobot. Palynol. 153, 70–85.
Gosling, W.D., Miller, C.S., Livingstone, D.A., 2013. Atlas of the tropical West African pollen flora. Rev. Palaeobot. Palynol. 199, 1–135.
Hall, J.B., Swaine, M.D., 1981. Distribution and ecology of vascular plants in a tropical rain forest. In: Junk, W. (Ed.), Forest Vegetation in Ghana. The Hague.
Havinga, A.J., 1964. Investigation into the differential corrosion susceptibility of pollen and spores. Pollen Spores VI, 621–635.
Havinga, A.J., 1984. A 20-year experimental investigation into the differential corrosion susceptibility of pollen and spores in various soil types. Pollen Spores XXVI, 541–558.
Holt, K., Allen, G., Hodgson, R., Marsland, S., Flenley, J., 2011. Progress towards an automated trainable pollen location and classifier system for use in the palynology laboratory. Rev. Palaeobot. Palynol. 167, 175–183.
Hooghiemstra, H., 1984. Vegetational and climatic history of the high plain of Bogotá, Colombia: A continuous record of the last 3.5 million years. Diss. Bot. 79, 1–138.
Hooghiemstra, H., Cleef, A.M., 1995. Pleistocene climatic change and environmental and generic dynamics in the north Andean montane forest and Paramo. In: Churchill, S. P., Balslev, H., Forero, E., Luyeyn, J.L. (Eds.), Biodiversity and Conservation of Neotropical Montane Forests. New York Botanical Garden (1993: Proc. symposium).
Jackson, S.T., 2012. Representation of flora and vegetation in Quaternary fossil assemblages: Known and unknown knowns and unknowns. Quat. Sci. Rev. 49, 1–15.
Jantz, N., Homeier, J., Behling, H., 2014. Representativeness of tree diversity in the modern pollen rain of Andean montane forests. J. Veg. Sci. 25, 481–490.
Koeberl, C., Milkereit, B., Overpeck, J.T., Scholz, C.A., Amoako, P.Y.O., Boamah, D., Danuor, S. K., Karp, T., Kueck, J., Hecky, R.E., King, J.W., Peack, J.A., 2007. An international and multidisciplinary drilling project into a young complex impact structure: the 2004 ICDP Bosumtwi Crater Drilling Project — an overview. Meteorit. Planet. Sci. 42, 483–511.
Krebs, C.J., 1999. Chapter 12 — Species Diversity Measures, Ecological Methodology, 2nd ed. Addison-Wesley Educational Publishers Inc.
Liu, K.-B., Colinvaux, P.A., 1985. Forest changes in the Amazon Basin during the last glacial maximum. Nature 318, 556–557.
Maher, L.J., 1972. Nomograms for computing 0.95 confidence limits of pollen data. Rev. Palaeobot. Palynol. 13, 85–93.
Miller, C.S., Gosling, W.D., 2014. Quaternary forest associations in lowland tropical West Africa. Quat. Sci. Rev. 84, 7–25.
Moore, P.D., Webb, J.A., Collinson, M.E., 1991. Pollen Analysis, Blackwell Scientific, Oxford.
Odgaard, B.V., 1999. Fossil pollen as a record of past biodiversity. J. Biogeogr. 26, 7–17.
Odgaard, B.V., 2001. Palaeoecological perspectives on pattern and process in plant diversity and distribution adjustments: a comment on recent developments. Divers. Distrib. 7, 197–201.
Reille, M., 1995. Pollen et spores d'Europe et d'Afrique du Nord, Laboratoire de Botanique Historique et Palynologie.
Roubik, D.W., Moreno, J.E.P., 1991. Pollen and Spores of Barro Colorado Island, Missouri Botanical Garden, United States.
Rull, V., 1987. A note on pollen counting in palaeoecology. Pollen Spores XXIX, 471–480.
Rull, V., Abbott, M.B., Polissar, P.J., Wolfe, A.P., Bezada, M., Bradley, R.S., 2005. 15,000-yr pollen record of vegetation change in the high altitude tropical Andes at Laguna Verde Alta, Venezuela. Quat. Res. 64, 308–317.
Shanahan, T.M., Overpeck, J.T., Beck, J.W., Wheeler, C.W., Peck, J.A., King, J.W., Scholz, C.A., 2008. The formation of biogeochemical laminations in Lake Bosumtwi, Ghana, and their usefulness as indicators of past environmental changes. J. Paleolimnol. 40, 339–355.
Smith, B., Wilson, J.B., 1996. A consumer's guide to evenness indices. Oikos 76, 70–82.
Stockmarr, J., 1971. Tablets with spores used in absolute pollen analysis. Pollen Spores 13, 615–621.
Sugita, S., 1994. Pollen representation of vegetation in Quaternary sediments: theory and method in patchy vegetation. J. Ecol. 82, 881–897.
Valencia, B.G., Urrego, D.H., Silman, M.R., Bush, M.B., 2010. From ice age to modern: a record of landscape change in an Andean cloud forest. J. Biogeogr. 37, 1637–1647.
Van der Knaap, W.O., 2009. Estimating pollen diversity from pollen accumulation rates: A method to assess taxonomic richness in the landscape. The Holocene 19, 159–163.
Van Der Knaap, W.O., Van Leeuwen, J.F.N., Froyd, C.A., Willis, K.J., 2012. Detecting the provenance of Galápagos non-native pollen: the role of humans and air currents as transport mechanisms. The Holocene 22, 1373–1383.
Vincens, A., Lézine, A.-M., Buchet, G., Lewden, D., Le Thomas, A., 2007. African pollen database inventory of tree and shrub pollen types. Review of Palaeobotany and Palynology 145, 135–141.
Williams, J.J., Gosling, W.D., Brooks, S.J., Coe, A.L., Xu, S., 2011a. Vegetation, climate and fire in the eastern Andes (Bolivia) during the last 18,000 years. Palaeogeogr. Palaeoclimatol. Palaeoecol. 312, 115–126.
Williams, J.J., Gosling, W.D., Coe, A.L., Brooks, S.J., Gulliver, P., 2011b. Four thousand years of environmental change and human activity in the Cochabamba Basin, Bolivia. Quat. Res. 76, 58–68.