Novel methods of estimating Relative Pollen Productivity, a key parameter for reconstruction of past land cover from pollen records

Yiman Fang [a], Chunmei Ma [b,c], M. Jane Bunting [a]*

School of Environmental Sciences, Cohen Building, University of Hull, Cottingham Road, Hull HU6 7RX, UK

School of Geographic and Oceanographic Sciences, Nanjing University, Nanjing 210023, China

Jiangsu Collaborative Innovation Center for Climate Change, China

* Corresponding author: e-mail: m.j.bunting@hull.ac.uk telephone: (+) 1482 466068

ABSTRACT

Reconstructing land cover from pollen data using mathematical models of the relationship between them has the potential to translate the many thousand pollen records produced over the last 100 years (over 2300 radiocarbon-dated pollen records exist for the UK alone – M. Grant pers comm) into formats relevant to ecologists, archaeologists and climate scientists. However, the reliability of these reconstructions depend on model parameters. A key parameter is Relative Pollen Productivity (RPP), usually estimated from empirical data using "Extended R Value analysis" (ERV analysis). Lack of RPP estimates for many regions is currently a major limitation on reconstructing global land cover.

We present two alternatives to ERV analysis, the Modified Davis method and an Iteration method, which use the same underlying model of the relationship between pollen and vegetation to estimate RPP from empirical data, but with different assumptions. We test them in simulation against ERV analysis, and use a case study of a problematic empirical dataset to determine whether they have the potential to increase the speed and geographic range of RPP estimation. The two alternative methods are shown to perform at least as well as ERV analysis in simulation. We also present new RPP estimates from southeastern sub-tropical China for 9 taxa estimated using the Modified Davis Method.

Adding these two methods to the "toolkit" for land cover reconstruction from pollen records opens up the possibility to estimate a key parameter from existing datasets with less field time than using current methods. This can both speed up the inclusion of more of the globe in past land cover mapping exercises such as the PAGES Landcover6k working group and improve our understanding of how this parameter varies within a single taxon and the factors control that variation.

Key-words:

Extended R-value method (ERV), Iteration Method (IM), Modified Davis Method (MDM), Pollen analysis, reconstruction of landcover from pollen data, southeast China

1. INTRODUCTION

Land cover is a fundamental earth system variable, with well understood impacts on regional and global climate and hydrological cycles. Reconstructing past land cover is therefore an important step to improve testing of regional, continental and global scale climate models against known past climate conditions, and recent developments coordinated by the PAGES Working Group LandCover6k (http://www.pastglobalchanges.org/ini/wg/landcover6k/intro) using models of pollen dispersal and deposition combined with data from pollen records preserved in lake and mire sedimentary systems offer a substantial step forward in producing global land cover maps (Hellman et al., 2008a, b; Soepboer et al., 2010; Trondman et al., 2015; Bunting et al., 2018). However, extending this approach from core researched areas (e.g. North-west Europe, Mazier et al., 2012; temperate China, Li, 2016) to the rest of the globe is limited by the need to parameterise models for the main plant taxa present in each region, which is currently achieved using the "Extended R value" approach (hereafter ERV approach). This method was developed nearly 40 years ago (Parsons and Prentice, 1981; Prentice and Parsons, 1983; Sugita, 1994), requires large input datasets, and sometimes produces erratic results where the feasible sampling strategy or innate structure of the modern land cover mosaic deviates from the underlying assumptions.

The relationship between land cover and pollen dispersal and deposition (d&d) is assumed to have a generally linear form, which is usually written as:

$$y_{ik} = \alpha_i x_{ik} \qquad \text{(Equation 1)}$$

Algebraic terms are defined in table 1. In order to use this equation to translate pollen data into land cover estimates, the parameter α_i and any parameters needed to calculate x_{ik} need to be specified. Underlying this model is the assumption that α_i is a constant for a given region, and that changes in y_{ik} over time occur in response to changes in x_{ik}. The assumption that pollen productivity is a constant is problematic, since most plant species show some plasticity in reproductive allocation in response to environmental variation (e.g. changes in the number of flowers or seeds produced per unit of plant), and pollen trapping studies show interannual variations in pollen influx in the absence of changes in plant cover. These known variations are treated as "noise" in terms of the pollen productivity measure needed to interpret sedimentary pollen records, where individual pollen assemblages come from samples which amalgamate multiple years, and reconstructions typically cover "time windows" of 100-500 years and are based on multiple pollen samples within each window.

Pollen productivity is usually estimated empirically using modern samples, for which both pollen assemblage present and the surrounding vegetation from which the pollen came can be measured directly. Measuring absolute pollen influx is not always possible, depending on the methods used, but pollen percentage data are widely available. Land cover data (e.g. community composition) is also often only available in relative units. This creates a problem, because in pollen percentage data taxa cannot be considered independently – changes in the proportion of one taxon affect all other taxa, even if their influx remains constant. This can lead to apparent variations in pollen productivity, but

the ratio of pollen productivities of two taxa will remain constant (Davis, 1963). Therefore pollen productivity is usually estimated by setting one taxon as a reference taxon with pollen productivity of 1, and expressing the pollen productivity of all other taxa as a ratio with this one, that is, as Relative Pollen Productivities. In the context of empirical estimation, therefore, the model of the underlying relationship (Equation 1) becomes:

$$p_{ik} = (R_j i)v_{ik}$$ (Equation 2 – terms in table 1)

The ERV approach assumes that a study region has homogenous vegetation composition at a large scale (e.g. that any 10km x10km block within the study region has the same species composition as any other). Where this is the case, as the area of vegetation surveyed for the calculation of x_{ik} increases, the vegetation abundance of taxon i tends towards a constant, the overall abundance in the study region, and equation (1) can be rewritten as:

$$y_{ik} = \alpha_i x_{ik} + \omega_i$$ (Equation 3 – terms in table 1)

The minimum radius/area of vegetation at which the relationship between pollen and vegetation does not improve with the addition of another increment of vegetation data is termed the Relevant Source Area of Pollen (RSAP: Sugita, 1994). For empirical estimation of α_i using the ERV approach, vegetation survey is designed to cover an area typically around 2-5 times the anticipated RSAP (see e.g. Bunting et al., 2013), usually via a combination of direct field survey and remote sensing (e.g. aerial photograph interpretation). The ERV approach estimates the two constants α_i and ω_i for all taxa in a dataset simultaneously, using an iterative maximum likelihood strategy (Parsons and Prentice, 1981; Prentice and Parsons, 1983; Sugita, 1993).

In this paper, we present two alternatives to the ERV approach which use Equation 1 rather than Equation 2, taking advantage of technological advances in both increased computer speed and availability of remote sensed land cover data in the decades since the ERV approach was proposed. We hypothesize that these methods will be better able to return robust estimates of RPP than the ERV approach for small datasets and in landscapes where modern land cover is heterogenous at the scale of 10-100km (e.g. where the assumption of homogenous regional vegetation does not hold). We first test this in simulation, where the input RPP values are known, then demonstrate the application of all three methods to a dataset from a sub-tropical upland area in southeast China where no RPP values have yet been reported.

Table 1. Algebraic terms used in equations in the text

Algebraic term	Symbols	Definition
α_i	Lower case Greek letter alpha	Pollen productivity of taxon i (assumed to be a constant for a given study region)
d	Lower case Roman letter d	Radial distance from deposition point k within which vegetation composition has been recorded and distance-weighted, then summed
i	Lower case Roman letter i	A specific taxon
j	Lower case Roman letter j	The taxon selected as the reference taxon, with Relative Pollen Productivity set to 1.
k	Lower case Roman letter k	A specific site within a study region
p_{ik}	Lower case Roman letter p	Pollen percentage of taxon i at site k
a_i	Lower case Roman letter a	Empirically estimated pollen productivity of taxon i
$R_j i$	Upper case Roman letter R	Ratio between pollen productivity of type i and pollen productivity of type j (Relative Pollen Productivity of type i where type j is the reference taxon)
v_{ik}	Lower case Roman letter v	Distance weighted plant abundance (measured in relative units) of taxon i around site k
V_{ik}	Upper case Roman letter V	Non-distance-weighted vegetation abundance of taxon i around site k.
y	Lower case Roman letter y	Pollen influx
x_{ik}	Lower case Roman letter x	distance-weighted plant abundance (measured in absolute units) of taxon i relative to site k, summed over a specified radial distance from site k (d). Many different distance-weighting models can be used.

y_{ik}	Lower case Roman letter y	pollen influx from taxon i at site k
ω_i	Lower case Greek letter omega	the background pollen contribution from taxon i in the study region, assumed to be a constant for taxon i in a given study region assuming that d extends to the Relevant Source Area of Pollen (Sugita, 1994)

2. METHODS

All analyses are applied to an input dataset where each sample consists of paired measurements of pollen proportions and distance weighted plant abundance (dwpa) for a common list of taxa recorded for multiple sites within a study area, either from simulation or from direct data collection. For ERV analysis, a widely used "rule of thumb" for minimum number of samples is 2n, twice the number of taxa being studied (Sugita, pers. comm.; Bunting et al., 2013), with more samples recommended. Input datasets were inspected using scatter plots of pollen percentage against dwpa summed to 2000m. A reference taxon was selected on grounds of having a wide range of both pollen percentages and dwpa values and showing a broadly linear scatter, and taxa with few non-zero data points or multiple outliers excluded (see e.g. Bunting et al., 2016).

2.1 The ERV Approach

ERV analysis was carried out using PolERV (Middleton, unpublished), which is a user-friendly wrapper around the ERV analysis code developed by Shinya Sugita (multiple versions), as described in more detail in Bunting and Hjelle (2010). ERV analysis has three variants, referred to as sub-models, each differing in terms of the assumption made about the definition of the background pollen component (e.g. Sugita, 1994); all three were used in these analyses.

2.2 The Modified Davis Method (MDM)

Davis (1963) defined the "R-value", a measure of pollen productivity, as the ratio between pollen proportion and vegetation abundance, and argued that the ratio of the R-values for a pair of taxa should be constant between sites. In this seminal paper, vegetation is expressed as area of cover with no distance weighting applied, and pollen input from beyond the surveyed area (background pollen) is assumed to be negligible. We therefore modify Davis' approach by using distance weighted vegetation data collected to a distance many times larger than the likely RSAP, to include most of the possible sources of background pollen input. The calculation can be written as:

$$R_j i = \frac{p_{ik}}{v_{ik}} . \frac{v_{jk}}{p_{jk}}$$ (Equation 4)

A more detailed explanation of the derivation of this equation can be found in Appendix 1. MDM calculations were carried out in Excel for each sample in the dataset independently.

2.3 The Iteration Method (IM)

The simplifying assumption of constant background in equation 3 meant that vegetation survey areas were manageable, but also required two parameters to be estimated for each taxon, which led to the development of the ERV method (Prentice and Parsons, 1981; Parsons and Prentice, 1983). When dwpa can be calculated for an area much larger than the RSAP, equation 2 can be used for all taxa, and a much simpler iteration approach developed since only one parameter per taxon is now required. We take a possible set of RPP values for the taxa, calculate the estimated pollen proportions from the known dwpa values using equation 2, then compare these pollen assemblages with the actual assemblages from the empirical dataset using summed squared chord distance (SSCD) as a measure of similarity. By trying many different combinations of RPP values, we can identify the set or sets with the lowest SSCD to identify which combination is the best estimate of the actual values. The calculations were carried out in R (code in Appendix 2).

2.4 Simulation studies

First, all three methods were compared in simulation, where the RPP values for the taxa can be defined by the user on the basis of assumptions about the known behavior of related taxa in other contexts and vegetation survey can be comprehensive. The same pollen dispersal and deposition model is used in the simulation to create a dataset of samples of paired vegetation survey and pollen count values which can then be used to estimate the RPP values. ERV analysis assumes homogeneity in the wider landscape, therefore we designed an experiment to compare homogenous and heterogenous landscapes.

The simulation study was carried out using HUMPOL0 software (Middleton, unpublished; earlier version published in Bunting and Middleton, 2005). Two land cover scenarios were created using Mosaic v3.2 (Middleton and Bunting, 2004; see Figure 1). Each scenario consisted of two grids, an outer one (20km x 20km, 20m pixels) and an inner one (5km x 5km, 10m pixels). In the first case (Figure 1a), both grids represented a landscape with a patchy vegetation structure which was homogenous (i.e. average vegetation within an area of 100 ha is the same throughout the landscape), and in the second case (Figure 1b) both grids represented a landscape with a patchy vegetation structure which was heterogenous (i.e. average vegetation composition in an area of 100ha varies depending on position in the landscape due to the presence of three distinct vegetation communities). The local patchiness of vegetation was created by distributing patches of six pollen taxa grouped into five plant communities in the landscape (community properties are given in Appendix 3). Three replicates of the inner grids (with different distributions of patches within the larger communities) allowed for collection of multiple sample points without risk of autocorrelation between points. Nine sample points were located in each inner grid, at least 1500m apart and 1000m from the boundary with the outer grid, giving 27 sample points in total for each dataset, which exceeds the 2n rule of thumb (six taxa, minimum of 12 samples).

[insert Figure 1 here]

From each sample point, simulated pollen loadings along with vegetation data collected in concentric 20m wide rings around the sampling point to a distance of 8km were calculated. Vegetation data were then distance weighted in Excel, using the Prentice-Sugita Gaussian Plume Model (Sutton, 1953; Prentice, 1985). Pollen assemblages are simulated as exact pollen loadings. In order to incorporate the sampling errors inherent in actual pollen counts, the loadings were used as a probability distribution to simulate pollen counts of 500 grains for each sample using Excel, and the counts were then expressed as percentages for further analysis.

2.5 Empirical data

After demonstrating the three methods in simulation, we applied them to a dataset from an upland area in southeastern China, where no RPP estimates are currently available. The available dataset consists of ten samples, and the landscape and sample set do not meet all the assumptions of ERV analysis (the wider landscape is heterogeneous, and samples cover habitats with different dominant species with few from mixed environments, meaning that in many cases values of dwpa are clumped rather than spread across a range). The dataset took 20 field days to collect, due to multiple logistical challenges typical of tropical environments, and serves as a good example of the issues faced in deriving RPP estimates for varied landscapes, and therefore being able to apply modern land cover reconstruction methods over much of the globe.

Figure 2 shows the chosen field area, the Meiling Mountains in Jiangxi province. The highest peaks reach about 950m a.s.l. (above sea level). The mountains lie along a southwest to northeast axis and occupy an area of 150 km^2 (28°31′N - 28°54′N, 115°34′E -115°53′E). The mountains are mainly composed of granite and gneiss. The mean annual temperature and total annual precipitation are about 12℃ and 1770mm, respectively. Due to human impact, the original vegetation has been largely replaced by secondary forest, with some primary forest remnants scattered in the area (Figure 2c). The main vegetation communities present are subtropical forests, including needleleaf forest (dominated by *Pinus massoniana* or *Cunninghamia lanceolata*), broadleaf deciduous forest (characterized by *Castanea sequinii*, *Quercus serrata* var. *breviptiolata* and *Platycarya strobilacea*), broadleaf evergreen forest (dominated by *Castanopsis sclerophylla* and *Cyclobalanopsis glauca*) and bamboo forest (*Phyllostachys edulis*). Shrubs are mainly distributed in the valleys, which support unstable communities caused by deforestation. Grassland are mainly found beside the small settlements of the mountains and on ridges with barren lands, and are dominated by Poaceae spp..

[Insert Figure 2 here]

10 sample points covering the main forest regions (described in Appendix 4) were chosen using a stratified random methodology. The following criteria were used to select RPP samples for empirical study: 1) a range of moderately accessible study locations were chosen to include samples with low, medium and high abundance in the plants of the main taxa 2) sample points were located randomly within the chosen locations, but the exact centre point was determined by practical limitations on the ground such as the availability of mosses to sample; 3) sample points were separated from each other by at least 200m;. The samples were collected from three broad vegetation communities, samples 1 and 4 in the *Cryptomeria - Cunninghamia - Phyllostachys* dominated forest, samples 2, 5 and 8 in the *Pinus* - Theaceae dominated forest, and samples 3, 6, 7, 9, 10 in the *Pinus - Cunninghamia - Cyclobalanopsis* forest. Vegetation survey around each point was conducted using the Crackles Bequest Project methodology (Bunting et al., 2013) via field mapping to 100m radius and supervised classification of Sentinel 2 imagery (ESA DATE data) for the larger region. Moss polsters were collected from each sampling point using inverted sample containers (Bunting et al., 2013) and prepared for pollen analysis using standard methods (addition of a *Lycopodium* spore tracer, treatment with 10% HCl, 10% KOH, HF and acetolysis mixture), then mounted in glycerine and identified under an optical microscope at ×400 magnification with reference to Wang et al. (1995), Tang et al. (2016) and photographs of pollen grains from a herbarium collected in Wuhan and Nanjing Botanical Gardens. Data were prepared for analysis using Survey (Middleton, unpublished; Farrell et al., 2016) and Excel, using the same dwpa weighting approach as described above. Pollen fall speeds (required for the distance weighting model) were calculated from measurements of multiple grains on sample slides using Stoke's Law, and are included in Appendix 5.

3. RESULTS

3.1 Simulation study

Figure 3 summarises the RPP estimates obtained in the simulation study from the two scenarios (homogenous or heterogenous landscape – Figure 1) with two different sample datasets (n=27 or n=9). *Cunninghamia* was used as the reference taxon. Since the actual RPP values for each taxon were specified by the user in the simulation, it is possible to compare the quality of the estimates obtained using the different methods. The rank order of RPP values is generally identified correctly by all methods, but the accuracy varies.

[insert Figure 3 here]

3.1.1 ERV analysis

The rank order of taxon RPP varies between scenarios when ERV analysis is used (see Figure 3), although the three groups of input values are always kept distinct. This apparent discrepancy partly arises because the error estimates provided by ERV analysis are small compared with the estimates used for the other two methods. The errors reported by the ERV analysis software are one standard deviation (SD) estimated by propagation, while the errors presented for the two new methods are the standard deviations of the full set of estimated RPP values, therefore it is possible to make finer discrimination of RPP rank.

Goodness of fit scores are produced by ERV analysis for each of additional vegetation data. These scores are plotted against distance in likelihood function plots (lf plots), which are then interpreted in terms of model behavior and robustness. An ideal plot falls monotonically to an asymptote at the RSAP distance. Figure 4 compares lf plots for the four simulation datasets analysed. None shows this ideal behaviour, but plots conform more closely to it for the big dataset in the homogeneous landscape (Figure 4a). The RSAP is identified using the "moving-window linear regression method" (Gaillard et al., 2008) using a 100m window. Estimates of RSAP from these plots range between 500m (SHo submodel 2 and 3) and 5000m (LHe submodel 1). Changing the homogeneity of the landscape and the size of the sample dataset has an effect on both the shape of ERV lf plots and the estimated RSAP.

[insert figure 4 here]

3.1.2 Modified Davis Method

MDM analysis always reproduces the input rank order of RPP values (see Table 1), albeit with fairly wide error margins. Unlike the other methods, which produce a single overall estimate of RPP, MDM analysis produces one estimate of RPP per sample. This allows the analyst to look at variation within the dataset, and potentially screen out samples with atypical behaviour or identify environmentally or taxonomically driven differences in RPP between sampled habitats. Figure 5 shows boxplots of the RPP estimates for all samples using MDM in the four scenarios, illustrating the presence of outliers even though the means are close to the input values.

[Insert Figure 5 here]

3.1.3 Iteration Method

In this study, we used a simplistic grid-search approach testing 6 possible values of RPP (0.1,0.5,1,2,4 and 10) for each taxon. In all four tests, the summed squared-chord distance (SSCD) is lowest when the pollen productivity of *Castanea* is 1, *Castanopsis* is 1, *Cunninghamia* is 1, *Cyclobalanopsis* is 0.5, *Pinus* is 2, and *Quercus* is 2, which are identical to the input RPPs values, but other solutions with low SSCD values occur with different RPP profiles. We therefore present the estimate of RPP as the mean of values for multiple good-fit solutions. The number of best-fit profiles for each test were selected individually based on their SSCD values, and expressed as an 'error bar' for each taxon. This method also returns the expected RPP estimates (see Figure 3 and Table 1).

3.2 Empirical study

Results from the empirical study are summarised in Figure 6.

[insert Figure 6 here]

Scatter plots (Figure 6a) compare pollen percent with dwpa for 9 taxa. *Pinus* was chosen as the reference taxon because it was present at all sites and had the widest range of values for both pollen percentage and dwpa. An underlying assumption of ERV analysis is that these scatter plots will not be linear, due to the interdependency of percentage data, and the ERV analysis process is designed to correct for these problems. The distance-weighting model chosen assumes that a single transport route, above canopy air flow, is the main control on pollen assemblage composition, and can be modelled as a linear relationship. However, in these scatter plots there appear to be two or more possible linear relationships. The high pollen/low dwpa points may represent cases where other transport mechanisms such as gravity or insects are dominating the pollen signal formation (Bunting et al., 2016).

Figure 6b shows the likelihood function scores from ERV analysis; the curves do not perform as expected (with values falling monotonically to an asymptote), probably reflecting the problems anticipated due to performing the analysis on a small dataset in a landscape which does not fit model assumptions well. The values do approach an asymptote, at around allowing estimates of RPP relative to *Pinus* (RPP_{Pinus}) to be

extracted. All three sub-models return similar results (Figure 6c), and the standard deviations for *Castanea*, *Quercus* and Rosaceae exceed the RPP_{Pinus} value, meaning that the RPP_{Pinus} for these three taxa is effectively zero. Poaceae has a very high RPP_{Pinus} value compared to all other taxa.

The RPP_{Pinus} values estimated using the modified Davis method are shown in figure 6d, with means shown in figure 6c as purple squares. The range of values is small for *Cryptomeria*, *Liquidambar*, Rosaceae, and Theaceae, whereas *Castanea* and *Quercus* which are both abundant in the pollen samples but rarely recorded in the vegetation survey have the largest variation. RPP_{Pinus} values are much smaller than 1 for most taxa, although very large outliers (>>50) for *Castanea* (5), *Cryptomeria* (2) and *Liquidambar* (1) were removed from the figure.

For the iteration method we chose to use a wider range of possible values than were used in the simulation study (0.01, 0.1, 0.25, 0.5, 1 and 4), and tested all combinations of these values for the 9 taxa. Results are summarised in figure 6c (green diamonds, with error estimates derived from the four combinations with the lowest SSCD values).

RPP_{Pinus} estimates from the three methods are broadly similar for most taxa, apart from Poaceae, *Castanea* and *Quercus*. Poaceae is the only herbaceous taxa included in the analysis, and the other two taxa have markedly clumped distribution in the vegetation and scatter plots.

4. Discussion

4.1 Simulation study

The simulation study showed that the two proposed alternative methods, modified Davis and Iteration, are at least as effective as the currently dominant Extended R Value analysis method at returning the input values of Relative Pollen Productivity, regardless of the number of sites included in the analysis. Therefore these two new methods are valid additions to the toolkit for land cover reconstruction from pollen data toolkit, and have the potential to allow analysts to more quickly produce reconstructions of land cover from areas which clearly do not meet the assumptions and requirements for effective ERV-analysis.

ERV analysis simulation results were not as strongly affected by smaller datasets or heterogeneous wider landscape as initially expected, apart from the difference in estimated RSAP, which can be explained by the difference in landscapes. The RSAP is less than 1500m in all cases of the homogeneous landscape, which accords with previous simulation studies using similar mosaic structures for vegetation (e.g. Bunting et al. 2004). However, the RSAP is close to 5000m in most cases with the heterogeneous landscape, reflecting the inclusion in the pollen signal of elements from all three different vegetation communities present; a larger RSAP reflects the larger area required to

encompass a homogenous mixture of the six plant types around each of the sampling points, regardless of which community they were actually located in.

None of the likelihood function score plots produced the "perfect" curve seen in other simulation studies (e.g. Hellman et al., 2009), probably due to a mixture of sample size, landscape scenario construction, sample location type and the introduction of pollen counting error effects into the simulation (in previous studies, pollen counts were assumed to be perfectly accurate - repeating the analysis multiple times with resampled pollen counts would produce a range of outputs and allow testing of the sensitivity of ERV analysis to counting effects). A larger number of samples (n=27) produces a smoother plot than the small dataset (n=9) in both landscapes, and the results from the homogenous landscape were closer to the ideal than those from the heterogenous landscape.

Both alternative methods require input vegetation data across a much larger area than the ERV analysis approach. The ERV method assumes that the pollen signal includes both a local component from within the surveyed area and a background pollen component from the wider landscape, and estimates the latter directly as part of the process. Both MDM and IM assume that the vegetation data presented encompass all possible sources of pollen for the sampled points (or, realistically, the substantial majority of those sources). The size of this vegetation survey area will increase rapidly with sampled basin size. With the wider availability of remote sensing data and computers capable of processing it, broad community maps are relatively straightforward to acquire, but the quality of the vegetation data input to the ERV analysis depends on the availability of appropriate ground-truthing data for community composition. The pollen dispersal and deposition model used assumes the information available is taxon canopy coverage (Bunting et al., 2013), and most available ground data (e.g. inventory plots, forestry data) does not measure this directly. If the analysis is carried out using a different vegetation metric, e.g. biomass or basal area, the results may not be comparable with RPP estimates derived from cover data, and resultant reconstructions need to be interpreted appropriately. Ideally, a planned field campaign to collect appropriate ground truthing data would be carried out, but in many cases this will not be immediately possible. Local data, the vegetation cover closest to the pollen sampling location (e.g. that within a 10-20m radius), will still need to be recorded in the field except when the sample is taken from a small lake or other location where there is no vegetation producing any of the pollen types of interest present within 10-20 meters of the sample point. An assumption must also be made about the extent of the vegetation area to include. Typically 40-60% of pollen at a sample point comes from within the RSAP (e.g. Sugita, 1994) but the long "tail" of the dispersal curve can extend substantially beyond it; where the sampling sites are clear canopy openings (e.g. mires in a wooded landscape, lakes) then the 70-90% characteristic radius of some pollen types can be 100s of kilometers (Prentice et al., 1987). Longer source areas are also inferred when different pollen dispersal and deposition models are used, such as the Lagrangian Stochastic Model (Theuerkauf et al., 2013).

We used a limited range of possible values in the iteration approach in this paper, coupled with a simple grid search methodology, and assessed difference with summed squared chord distances. Now that the concept is demonstrated, future developments will include using a more efficient sampler such as a variant on a Markov-Chain Monte Carlo method to allow a wider range of values to be tested without concomitant increases in run time. Whilst the SSCD is an unusual measure of fit for estimation studies considered broadly, it is widely used in palaeoecological research, and therefore chosen here partly for its familiarity to the target research community. As the work develops alternative measures of fit will be explored.

The Modified Davis Method obtains estimates of RPP from each sample individually, rather than a single estimate with associated errors from the whole dataset. This offers the potential of exploring location-related differences in RPP estimates, which may allow a more informed selection of outliers for removal from mean calculations, or investigation of the effects of differences in local conditions on RPP. For example, in the empirical study samples producing very high outlier values for RPP_{Pinus} of *Castanea* (>50) are mostly located close to *Castanea* trees, which suggests that pollen was being delivered to those moss pollsters by a localized transport mechanism such as gravity deposition, in addition to the above-canopy air flow transport assumed by the model.

4.2 Empirical Study

The empirical dataset used here has many of the characteristics of problematic datasets for ERV analysis. The sample locations were not fully randomly distributed (Broström et al., 2005; Mazier et al., 2008) and the range of vegetation cover values for some important taxa was very limited (Broström et al., 2008). The likelihood function curves did not behave as expected, but stable solutions were apparently found. Three of the nine taxa had unreliable RPP estimates, where the RPP estimate was smaller than the standard deviation (definition of unreliable RPP estimates follows e.g. Li et al., 2017). The results produced do not provide many, if any, usable estimates of RPP (see e.g. regional reviews by Mazier et al. 2012 and Li et al., 2018).

Estimation of RPP using the modified Davis method (Figure 6c) produced a wide range of values. The sample locations (Appendix 4) are in several different vegetation communities, which may partly explain the variations seen, and in some cases the scatter plots suggest that the taxa are distributed by transport mechanisms such as gravity or insect movement rather than solely by wind, therefore a model assumption is breached (see e.g. Bunting et al., 2016). For example, *Castanea* pollen was sometimes found on slides as clumps, suggesting that a whole anther had fallen into the moss sample from a nearby tree. As discussed above, examination of the individual values offered confirmatory evidence that in some cases *Castanea* pollen was being delivered to moss polsters by at least two routes, not just the single above canopy air flow route assumed by the pollen dispersal and deposition model used. Examination of the pollen-dwpa scatter plot for *Castanea* (Figure 6a) shows a non-linear pattern, similar to those seen for *Fraxinus* in larger dataset empirical studies by Bunting et al. (2016) which were interpreted as showing the influence of a second mode of pollen transport. Using the

MDM potentially allows for clear identification of problematic samples, and for a reasoned determination of RPP estimate – for reconstruction of land cover from pollen records from sediment cores from basins larger than c. 50m in radius, which will not be near any individual trees, an estimate taken from the mean of the lower group of values is more appropriate. For interpretation of records from small sites such as forest hollows, this MDM highlights an important factor which might be overlooked in the blanket application of quantitative methods.

Results are presented relative to *Pinus*, a taxon which is widely accepted to be a high pollen producer. In order to compare the RPPs with those obtained with other studies (Table 2), we recalculated the pollen productivity relative to Poaceae (Table 2). The results of the three methods we applied differ from each other substantially, but given the small dataset and individual site effects identified through using the MDM results, we argue that the MDM results with major outliers removed are the most appropriate for use in future land cover reconstruction. The MDM RPP value for *Castanea* (1.14) is an order of magnitude lower than that found in farmland by Li et al., (2017; 11.49). This might be due to the low abundance of *Castanea* in the Meiling Mountains, where the forest is dominated by economically valuable species such as *Pinus massoniana*, *Phyllostachys edulis* and *Cunninghamia lanceolata*, to ecological differences between *Castanea* pollen production and dispersal in the two areas due in particular to the different habitats where the pollen samples were collected, or to the accidental inclusion of some samples affected by gravity input of *Castanea* in the Li et al. (2017) dataset. The MDM RPP of *Pinus* is 5.6 in this study which is close to the mean value from Europe (6.38) but lower than published studies in northern and temperate China (Li et al., 2015; Li et al., 2017; Zhang et al., 2017), whilst the value for *Quercus* (9.14) is higher; this may reflect differences in species present in the two regions and/or differences in habitat between temperate and sub-tropical regions. Whilst a larger dataset to produce more reliable RPP estimates is desirable, using the Modified Davis Method allows us to present initial values which can be used for land cover reconstruction, whereas the ERV analysis method does not.

Table 2. Comparison of the relative pollen productivities (RPP) in this study with RPPs obtained in two other Chinese studies, and RPP values from Europe

Pollen type	This paper (ERV analysis)	This paper (modified Davis method*)	This paper (iteration method	Li et al. (2015) Woodland; Changbai Mountains	Li et al. (2017) cultural landscape, Shandong	Zhang et al., (2017) Changbai Mountains	Li et al., (2018) Temperate China**	Mazier et al. (2012) Europe
Castanea	0.02±0.15	1.14±0.36	0.25		11.49±0.49		11.49±0.49	
Cryptomeria	0.24±0.01	1.63±0.99	0.1					
Cyclobalanopsis	0.0002±0.0003	1.79±0.65	0.01					
Liquidambar	0.0535±0.01	1.39±1.04	0.01					
Pinus	0.28±0.01	5.6±4.3	4	16.13±0.52	8.96 ± 0.23	18.82±0.54	18.37±0.48	6.38 ± 0.45
Poaceae	1	1	1	1	1	1	1	1
Quercus	0.005±0.02	9.14±5.68	0.25	5.19±0.09	4.89 ±0.16	1.75±0.31	5.19±0.07	5.83 ± 0.15
Rosaceae	0.006±0.03	0.41±0.28	0.1					
Theaceae	0.001±0.01	0.11±0.07	0.01					

* Outliers removed

** Synthesis of RPP values

4.3 Implications for palaeoecology and future directions

Translation of pollen records into quantitative reconstructions of past land cover will open up a wealth of data on long term vegetation dynamics which has great potential for investigating key research questions in ecology, archaeology and climate change science. The PAGES Landcover6k working group, for example, is producing land cover maps which will feed into the PMIP simulations testing the performance of the global and regional climate models which are used to predict future climate change. Landcover6k reconstructs land cover using models of pollen dispersal and deposition which depend on estimates of Relative Pollen Productivity to function. Obtaining these estimates using the standard ERV analysis approach can be problematic and time consuming, and the methods proposed in this paper offer useful alternatives.

The methods presented here allow analysts to extract estimates of RPP from small datasets, or datasets that were collected using sampling strategies which do not meet the assumptions of ERV analysis. This includes a wide range of existing datasets such as the records from Tauber Traps collected by the Pollen Monitoring Programme (Hicks et al., 2001; http://www.pollentrapping.org/), or top samples from lake records, available in regions of the world which so far have no published RPP estimates. The Modified Davis Method, which calculates separate estimates from each sample, also has potential to support much-needed research into controls on observed variations in RPP in response to factors such as habitat, site management, or taxonomic variation in source plants.

Future investigations using these methods should include further testing in simulation to understand how vegetation mosaic structure affects their performance, to investigate method behaviour when used with different pollen dispersal and deposition models, and to explore and quantify the effects of counting error. Testing the effects of a larger suite of possible RPP values in the iteration method, and encoding the method more efficiently, will increase the usefulness of this approach. Development of vegetation survey strategies to determine the minimum field survey distance needed for vegetation map creation given the "pollen's eye view" of land cover (see Bunting et al., 2013) will improve the speed of RPP estimation using all three methods.

5. Conclusion

We have presented two new methods of estimating Relative Pollen Productivity from empirical datasets which are alternatives to the widely used ERV analysis approach, and demonstrated that, in simulation, these methods work at least as well as ERV analysis. We then applied these methods to a small dataset from sub-tropical southeast China, and are able to present a first set of RPP values for the region using the alternative methods although ERV analysis did not produce a useful solution. The RPP values obtained for pollen types which also occur in the better-studied temperate regions of eastern China (e.g. *Quercus*, *Pinus*) show clear differences, which we suggest may be due to differences in the species present, and recommend be tested further.

Acknowledgements.

Thanks are due to Lei Yang (Wuhan University, China), Jue Sun, Liang Li, Dr. Lin Zhao (Nanjing University, China), Dr. Yulian Jia, Chaohao Ling, Jun Luo, Yuanhui Chen, Chuan Chen and Shiping Li (Jiangxi Normal University, China) for helping with fieldwork and plant identification. We thank Michelle Farrell (Coventry University, UK) and two anonymous referees for their useful comments on an earlier version of the manuscript. The study is a contribution to the PAGES LandCover6k working group (http://www.pastglobalchanges.org/ini/wg/landcover6k/intro).
Surface pollen, vegetation and SENTINEL II data from the empirical study will be made available through the University of Hull data repository.

Funding statement

This research was carried out with the support of a student grant for Yiman Fang from the China Scholarship Council (CSC, Grant number 201506190128) and Dudley Stamp Memorial Award from Royal Geographical Society (with IBG), UK (Grants DSMA 25/16), funds from the National Key Research and Development Program (2016YFA0600501), the National Natural Science Foundation of China (NSFC, Grants 41671196). We have no conflicts of interest.

References

Broström A, Nielsen AB, Gaillard MJ, Hjelle K, Mazier F, Binney H, Bunting J, Fyfe R, Meltsov V, Poska A and Räsänen S (2008) Pollen productivity estimates of key European plant taxa for quantitative reconstruction of past vegetation: a review. *Vegetation history and archaeobotany*. 17(5), 461-478.

Broström A, Sugita S, Gaillard MJ and Pilesjö P (2005) Estimating the spatial scale of pollen dispersal in the cultural landscape of southern Sweden. *The Holocene*, 15(2), 252-262.

Bunting MJ and Middleton R (2005) Modelling pollen dispersal and deposition using HUMPOL software, including simulating windroses and irregular lakes. *Review of Palaeobotany and Palynology*, 134(3-4), 185-196.

Bunting MJ, Grant MJ and Waller M (2016) Approaches to quantitative reconstruction of woody vegetation in managed woodlands from pollen records. *Review of Palaeobotany and Palynology*, 225, 53-66.

Bunting MJ and Hjelle KL (2010) Effect of vegetation data collection strategies on estimates of relevant source area of pollen (RSAP) and relative pollen productivity estimates (relative PPE) for non-arboreal taxa. *Vegetation History and Archaeobotany*, 19(4), 365-374.

Bunting MJ, Farrell M, Bayliss A, Marshall P and Whittle A (2018) Maps from mud—using the multiple scenario approach to reconstruct land cover dynamics from pollen records: a case study of two Neolithic landscapes. *Frontiers in Ecology and Evolution*, 6, 36.

Bunting MJ, Farrell M, Broström A, Hjelle KL, Mazier F, Middleton R, Nielsen AB, Rushton E, Shaw H and Twiddle CL (2013) Palynological perspectives on vegetation survey: a critical step for model-based reconstruction of Quaternary land cover. *Quaternary Science Reviews*, 82, 41-55.

Davis MB (1963) On the theory of pollen analysis. *American Journal of Science*, 261(10), 897-912.

Eisenhut G (1961) *Untersuchungen über die Morphologie und Ökologie der Pollenkörner heimischer und fremdländischer Waldbäume*. Forstwissenschaftliche Forschungen (translated into English by Jackson,T.S. and Jaumann, P., 1989).

Farrell M, Bunting MJ and Middleton R (2016) Replicability of data collected for empirical estimation of relative pollen productivity. *Review of Palaeobotany and Palynology* 232 1-13.

Gaillard MJ, Sugita S, Bunting MJ, Middleton R, Broström A, Caseldine C, Giesecke T, Hellman SEV, Hicks S, Hjelle K, Langdon C, Nielsen AB, Poska A, von Stedingk H, Veski S and POLLANDCAL members (2008) The use of modelling and simulation approach in reconstructing past landscapes from fossil pollen data: a review and results from the POLLANDCAL network. *Vegetation History and Archaeobotany*, 17(5), 419-443.

Hellman S, Gaillard MJ, Broström A and Sugita S (2008a) The REVEALS model, a new tool to estimate past regional plant abundance from pollen data in large lakes: validation in southern Sweden. *Journal of Quaternary Science*, 23(1), 21-42.

Hellman SE, Gaillard MJ, Broström A and Sugita S (2008b) Effects of the sampling design and selection of parameter values on pollen-based quantitative reconstructions of

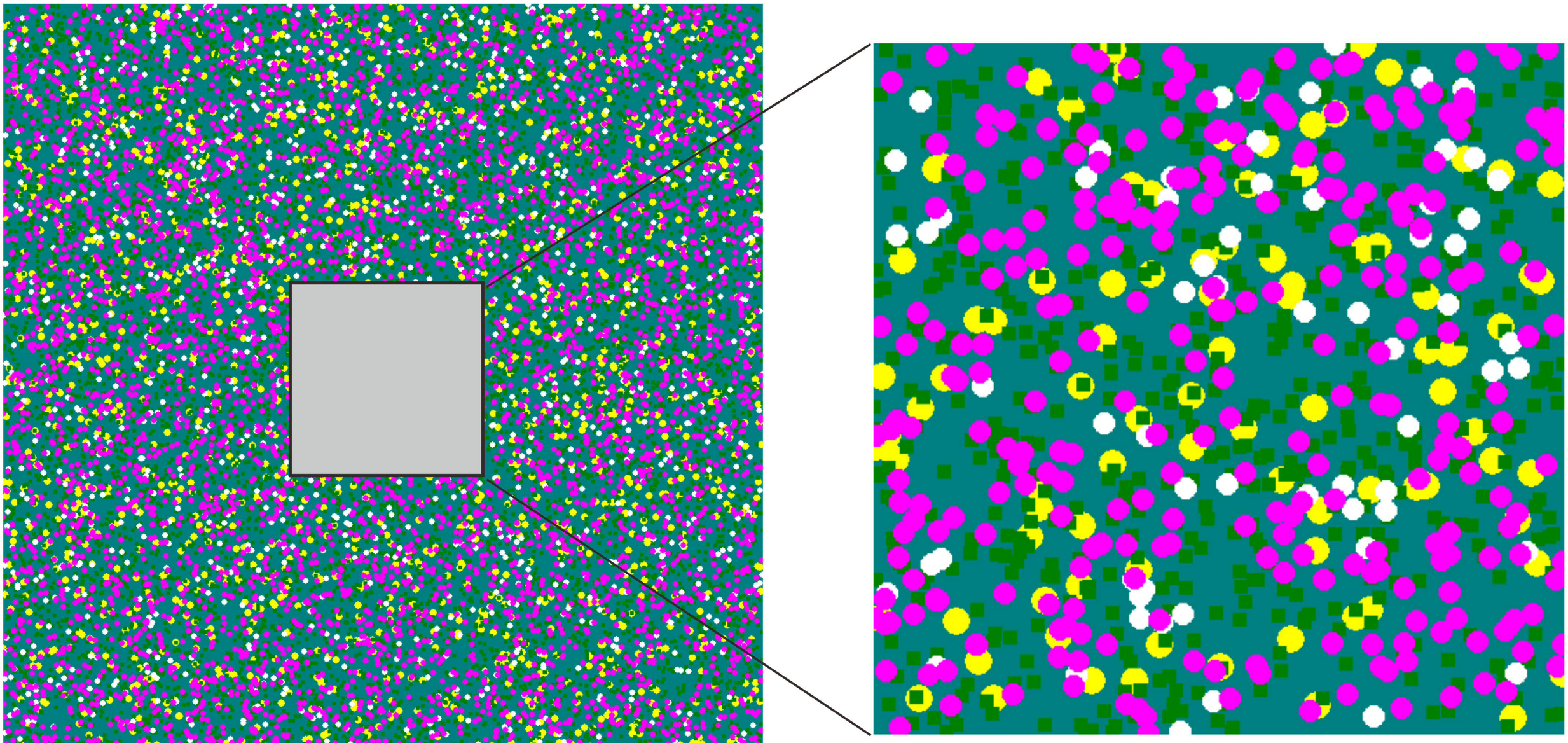
regional vegetation: a case study in southern Sweden using the REVEALS model. *Vegetation History and Archaeobotany*, 17(5), 445-459.
Hellman SE, Gaillard MJ, Bunting MJ and Mazier F (2009) Estimating the relevant source area of pollen in the past cultural landscapes of southern Sweden - a forward modelling approach. *Review of Palaeobotany and Palynology*, 153(3), 259-271.
Hicks S, Tinsley H, Huusko A, Jensen C, Hättestrand M, Gerasimides A and Kvavadze E (2001) Some comments on spatial variation in arboreal pollen deposition: first records from the Pollen Monitoring Programme (PMP). *Review of Palaeobotany and Palynology*, 117(1-3), 183-194.
Li F (2016) *Pollen productivity estimates and pollen-based reconstructions of Holocene vegetation cover in northern and temperate China for climate modelling*. PhD thesis, Linnaeus University, Kalmar.
Li F, Gaillard MJ, Sugita S, Mazier F, Xu Q, Zhou Z, Zhang Y, Li Y and Laffly D (2017) Relative pollen productivity estimates for major plant taxa of cultural landscapes in central eastern China. *Vegetation History and Archaeobotany*, 26(6), 587-605.
Li F, Gaillard MJ, Xu Q, Bunting MJ, Li Y, Li J, Mu H, Lu J, Zhang P, Zhang S, Cui Q, Zhang Y and Shen W (2018) A review of relative pollen productivity estimates from temperate China for pollen-based quantitative reconstruction of past plant cover. *Frontiers in plant science*, 9.
Li Y, Nielsen AB, Zhao X, Shan L, Wang S, Wu J and Zhou L (2015) Pollen production estimates (PPEs) and fall speeds for major tree taxa and relevant source areas of pollen (RSAP) in Changbai Mountain, northeastern China. *Review of Palaeobotany and Palynology*, 216, 92-100.
Mazier F, Brostöm A, Gaillard MJ, Sugita S, Vittoz P and Buttler A (2008). Pollen productivity estimates and Relevant Source Area for major taxa in a pasture woodland (Jura mountains, Switzerland). *Vegetation History and Archaeobotany*, 17, 479-495.
Mazier F, Gaillard MJ, Kuněs P, Sugita S, Trondman AK and Broström A (2012) Testing the effect of site selection and parameter setting on REVEALS-model estimates of plant abundance using the Czech Quaternary Palynological Database. *Review of Palaeobotany and Palynology*, 187, 38-49.
Middleton R and Bunting MJ (2004) Mosaic v1. 1: landscape scenario creation software for simulation of pollen dispersal and deposition. *Review of Palaeobotany and Palynology*, 132(1-2), 61-66.
Parsons RW and Prentice IC (1981) Statistical approaches to R-values and the pollen-vegetation relationship. *Review of Palaeobotany and Palynology*, 32(2), 127-152.
Prentice IC and Parsons RW (1983) Maximum likelihood linear calibration of pollen spectra in terms of forest composition. *Biometrics*, 1051-1057.
Prentice IC, Berglund BE, and Olsson T (1987) Quantitative forest - composition sensing characteristics of pollen samples from Swedish lakes. *Boreas*, 16(1), 43-54.
Prentice IC (1985) Pollen representation, source area, and basin size - toward a unified theory of pollen analysis. *Quaternary Research*, 23, 76–86.
Soepboer W, Sugita S and Lotter AF (2010) Regional vegetation-cover changes on the Swiss Plateau during the past two millennia: a pollen-based reconstruction using the REVEALS model. *Quaternary Science Reviews*, 29(3-4), 472-483.
Sugita S (1993) A model of pollen source area for an entire lake surface. *Quaternary Research*, 39, 239-244.

Sugita S (1994) Pollen representation of vegetation in Quaternary sediments: theory and method in patchy vegetation. *Journal of Ecology*, 881-897.
Sutton OG (1953) *Micrometeorology: a study of physical processes in the lowest layers of the earth's atmosphere* (No. 551.50 SUT). New York: McGraw-Hill.
Tang LY, Mao LM, Shu JW, Li ChH, Shen CM and Zhou ZhZ (2016) *An Illustrated Handbook of Quaternary Pollen and Spores in China.* Science Press, Beijing.
Trondman AK, Gaillard MJ, Mazier F, Sugita S, Fyfe R, Nielsen AB, Twiddle C, Barratt P, Birks HJB, Bjune AE, Björkman L, Broström A, Caseldine C, David R, Dodson J, Dörfler W, Fischer E, van Geel B, Giesecke T, Hultberg T, Kalnina L, Kangur M, van der Knaap P, Koff T, Kuneš P, Lagerås P, Latałowa M, Lechterbeck J, Leroyer C, Leydet M, Lindbladh M, Marquer L, Mitchell FJG, Odgaard BV, Peglar SM, Persson T, Poska A, Rösch M, Seppä H, Veski S and Wick L (2015) Pollen-based quantitative reconstructions of Holocene regional vegetation cover (plant-functional types and land-cover types) in Europe suitable for climate modelling. *Global Change Biology*, 21(2), 676-697
Theuerkauf M, Kuparinen A and Joosten H (2013) Pollen productivity estimates strongly depend on assumed pollen dispersal. *The Holocene*, 23(1), 14-24.
Wang FX, Qian NF and Zhang YL (1995) Pollen flora of China. Science Press, Beijing.
Zhang PP, Xu QH, Gaillard MJ, Mu HS, Zhang YH and Lu JY (2017) Research of main plant species's relative pollen productivities and relevant source area of temperate coniferous and broad-leaved mixed forest in Northern China. *Quaternary Science*, 37(6), 1429-1443

Figure captions

Figure 1. Landcover grids used for simulation of datasets for comparing models (outer grid: 20km x 20km; inner grid: 5km x 5km). a. homogeneous distribution of vegetation at the landscape scale, b. heterogeneous distribution of vegetation at the landscape scale
Figure 2. Location map and regional vegetation communities of the Meiling Mountains, southeast China (a. map showing the location of study area in China, b. map showing the Meiling Mountains in their wider landscape setting, c. map showing the inner study area and the locations of the 10 empirical sampling points (vegetation communities shown in the legend were mapped by classifying SENTINEL II data (European Space Agency (ESA) DATA) using field data from 2016 for ground truthing.
Figure 3. Estimates of relative pollen productivity obtained from the simulation study with different combinations of data set size and landscape distribution of vegetation. All values are relative to *Cunninghamia*. The open triangle shows the values input into the simulation. The black, grey and light grey circles show RPP for the three ERV submodels, estimated at the RSAP for each. The purple square shows RPPs estimated by the Modified Davis Method. The green diamond shows the mean of RPPs from the best-fit solutions estimated by the Iteration Method. The three letter code on each plot (e.g. LHo) identifies data set size (**L**arge or **S**mall) and landscape distribution of vegetation (**Ho**mogenous or **He**terogenous) for quick reference from the text. All error bars shown are one standard deviation.
Figure 4. Likelihood function score plot for the three ERV sub-models 1, 2 and 3 (a. large dataset (27 samples) in homogeneous landscape, b. large dataset (27 samples) in heterogeneous landscape, c. small dataset (9 samples) in homogeneous landscape, d. small dataset (9 samples) in heterogeneous landscape)
Figure 5. Box plots of individual sample estimates of RPP relative to *Cunninghamia* obtained by applying the Modified Davis Method to the four simulation scenarios.
Figure 6. Results from analysis of the empirical dataset from the Meiling Mountains. a) scatterplots of pollen percentage against dwpa to 2000m for nine key taxa b) Likelihood function scores from ERV analysis c) comparison of RPP estimates obtained using the three methods, with *Pinus* as the reference taxon and d) box plots of individual sample estimates of RPP obtained using the Modified Davis Method (values > 8 are not shown)

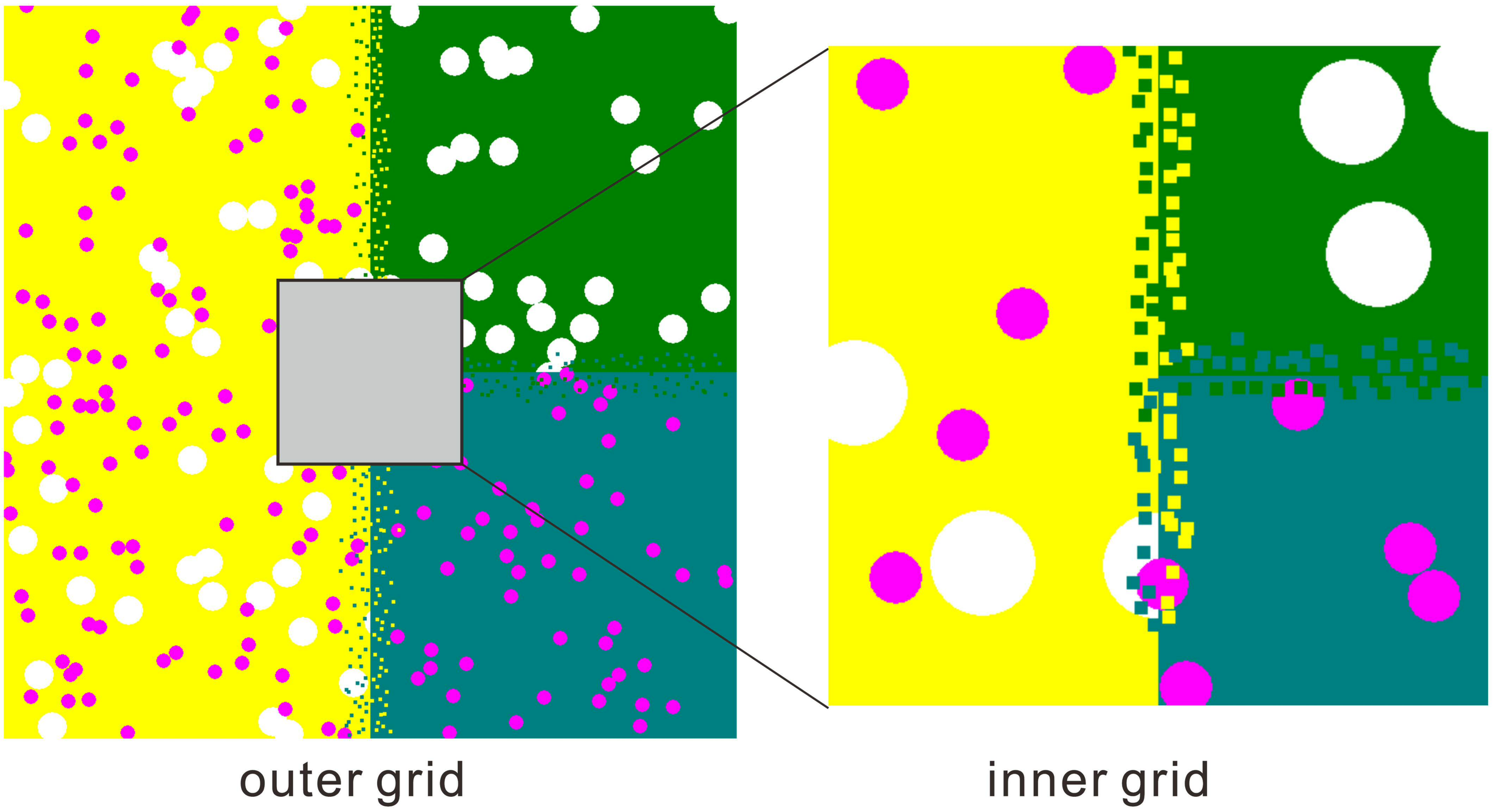
a



outer grid

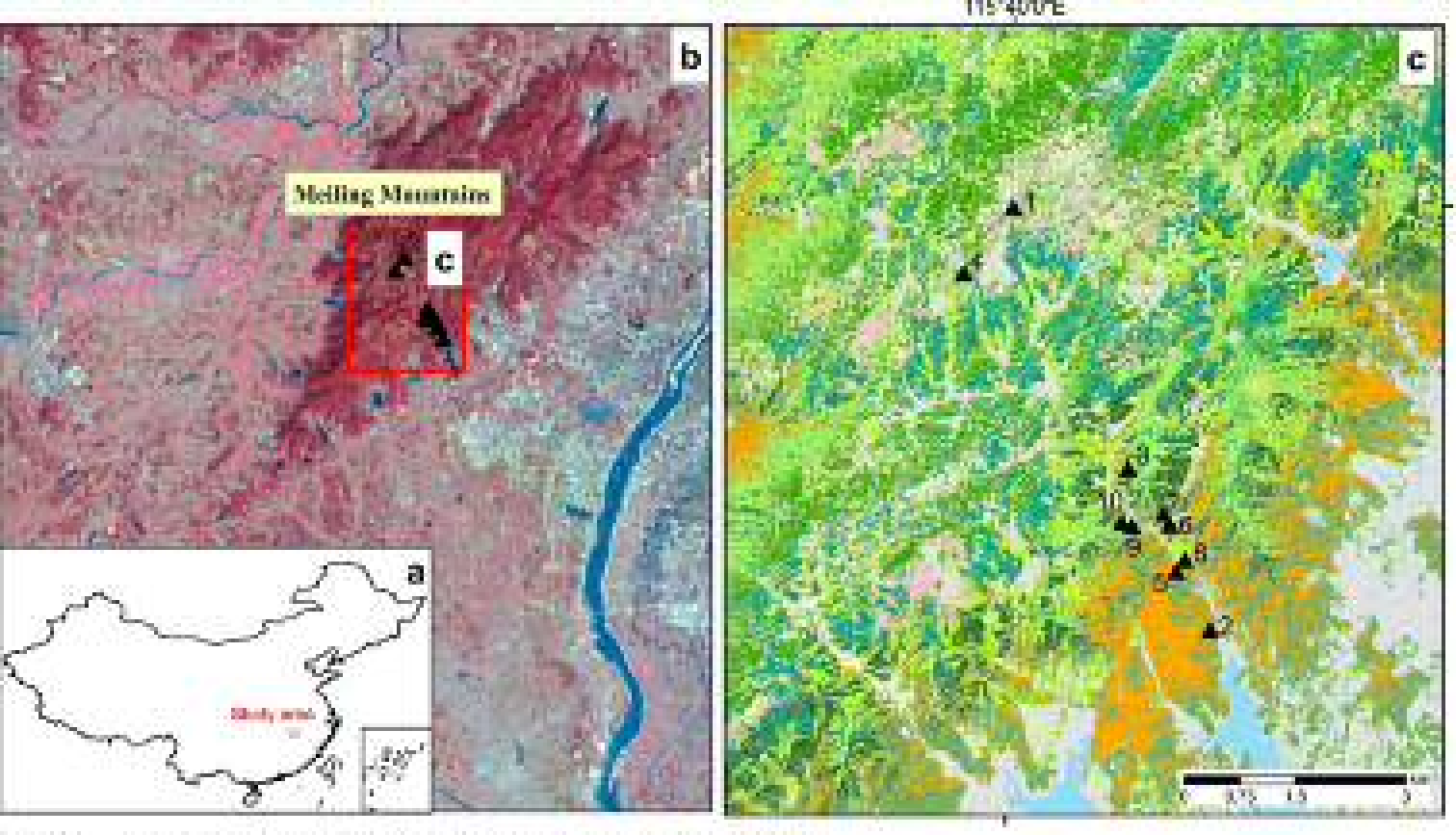
inner grid

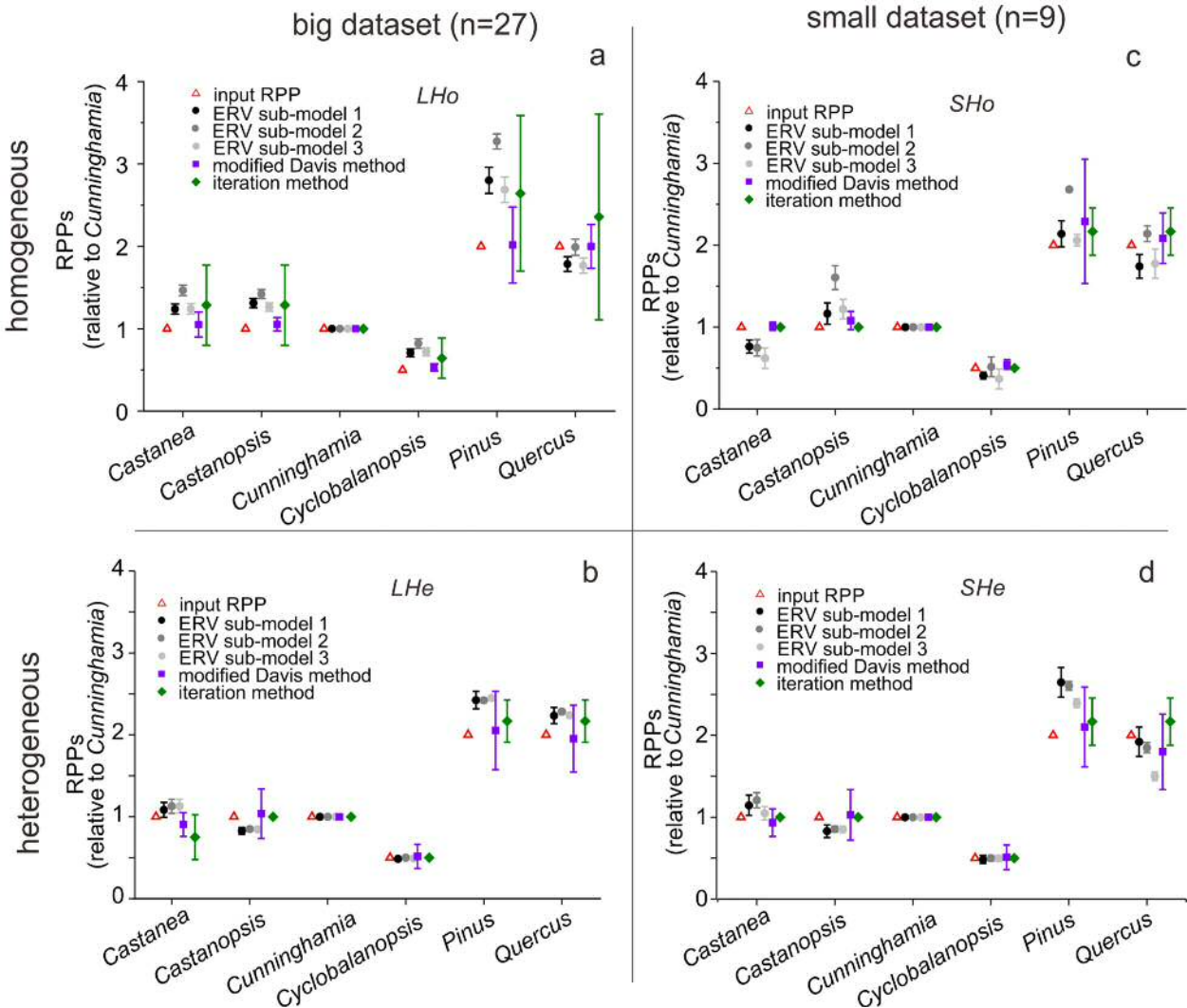
b



outer grid

inner grid


115°40'0"E
b
Meiling Mountain
c
a
Study area
c
Legend
13 RPP samples
(The vegetation communities only applied to part c of the figure)
water
Cryptomeria japonica var. sinensis forest
Phyllostachys edulis forest
Pinus massoniana - Cyclobalanopsis glauca - Liquidambar formosana mixed forest
Pinus massoniana - Cyclobalanopsis glauca - Liquidambar formosana mixed forest
Pleioblastus amarus - Arundinella anomala grassland
Pleioblastus amarus - Arundinella anomala grassland
Phyllostachys edulis forest
Cyclobalanopsis glauca - Pinus massoniana - Liquidambar formosana mixed forest
Poaceae spp. grassland


big dataset (n=27)
small dataset (n=9)
homogeneous
heterogeneous
a
b
c
d
LHo
LHe
SHo
SHe
input RPP
ERV sub-model 1
ERV sub-model 2
ERV sub-model 3
modified Davis method
iteration method
RPPs
(relative to Cunninghamia)
Castanea
Castanopsis
Cunninghamia
Cyclobalanopsis
Pinus
Quercus

homogeneous

heterogeneous

a

big dataset (n=27)

LHo

ERV Submodel 1
ERV Submodel 2
ERV Submodel 3

Likelihood function score

21100
21050
21000

0
2000
4000
6000
8000

Distance (m)

c

small dataset (n=9)

SHo

ERV Submodel 1
ERV Submodel 2
ERV Submodel 3

Likelihood function score

6690
6680
6670
6660
6650

0
2000
4000
6000
8000

Distance (m)

b

LHe

ERV Submodel 1
ERV Submodel 2
ERV Submodel 3

Likelihood function score

13650
13640
13630
13620
13610
13600
13590
13580

2000
4000
6000
8000

Distance (m)

d

SHe

ERV Submodel 1
ERV Submodel 2
ERV Submodel 3

Likelihood function score

4700
4690
4680

2000
4000
6000
8000

Distance (m)

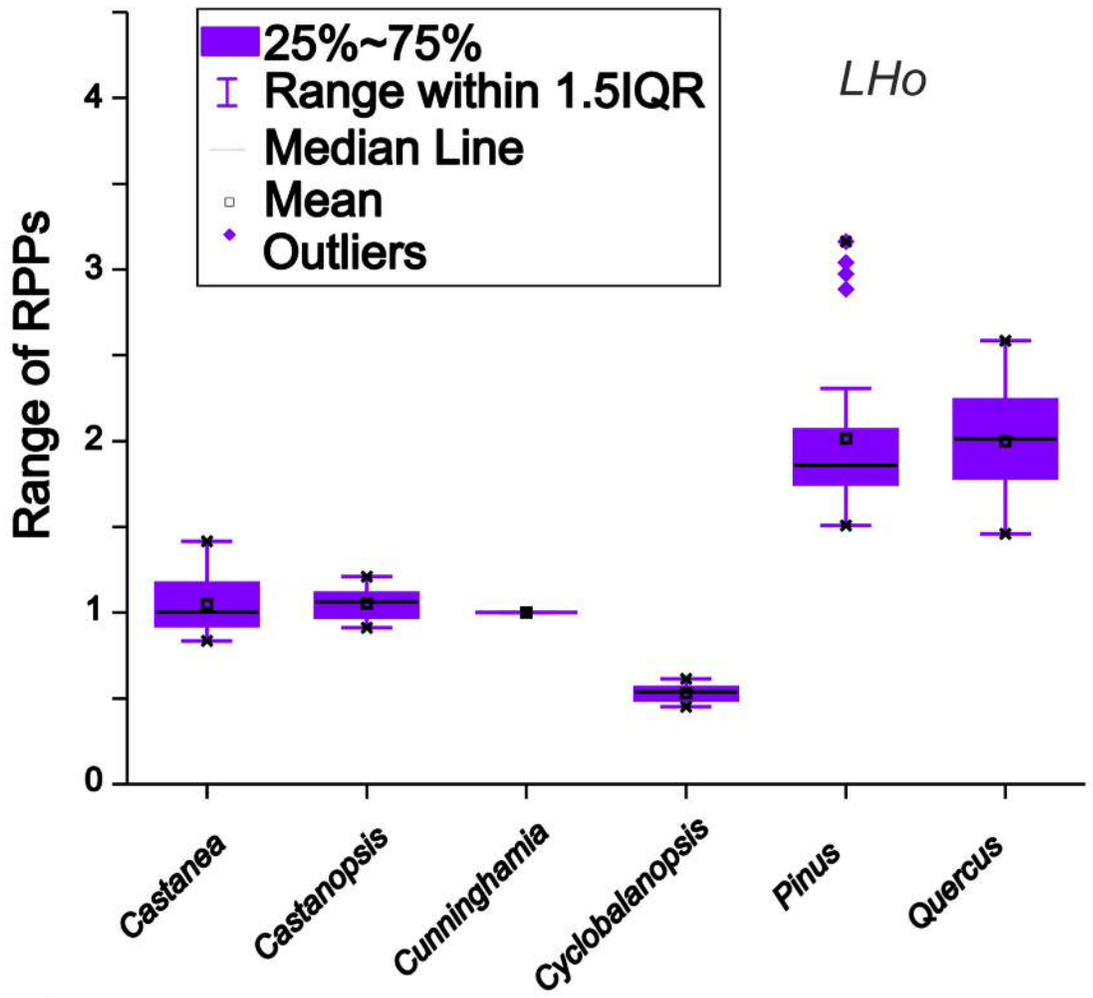
big dataset (n=27)

small dataset (n=9)

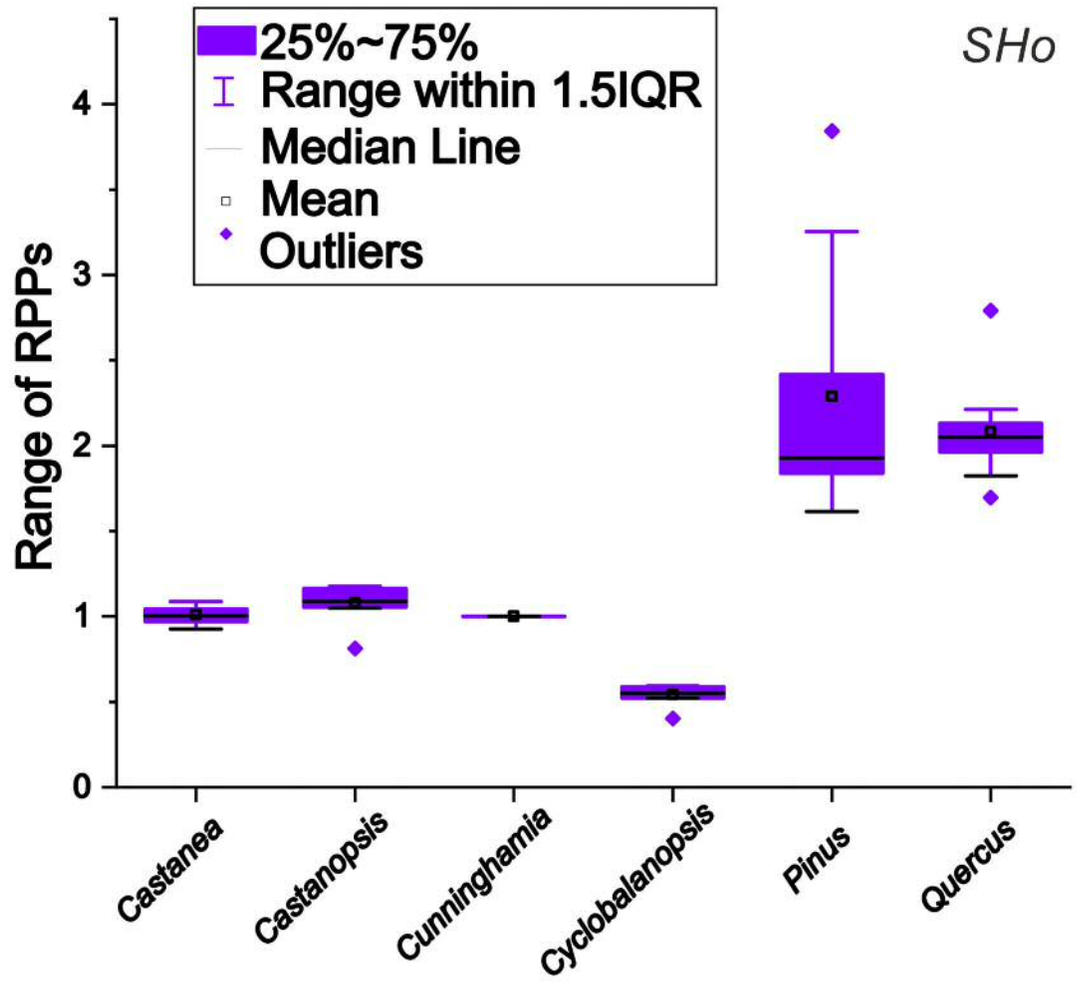
homogeneous

heterogeneous

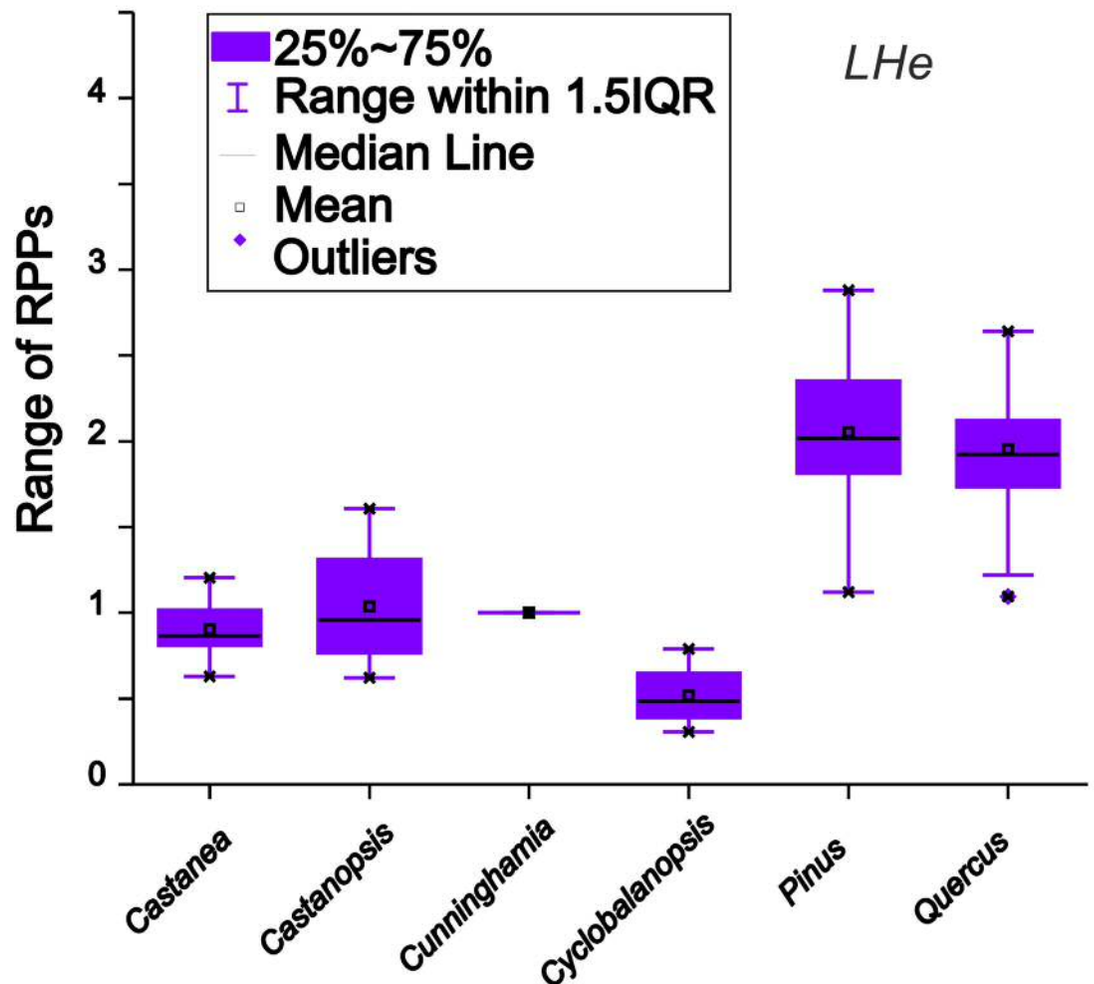
a



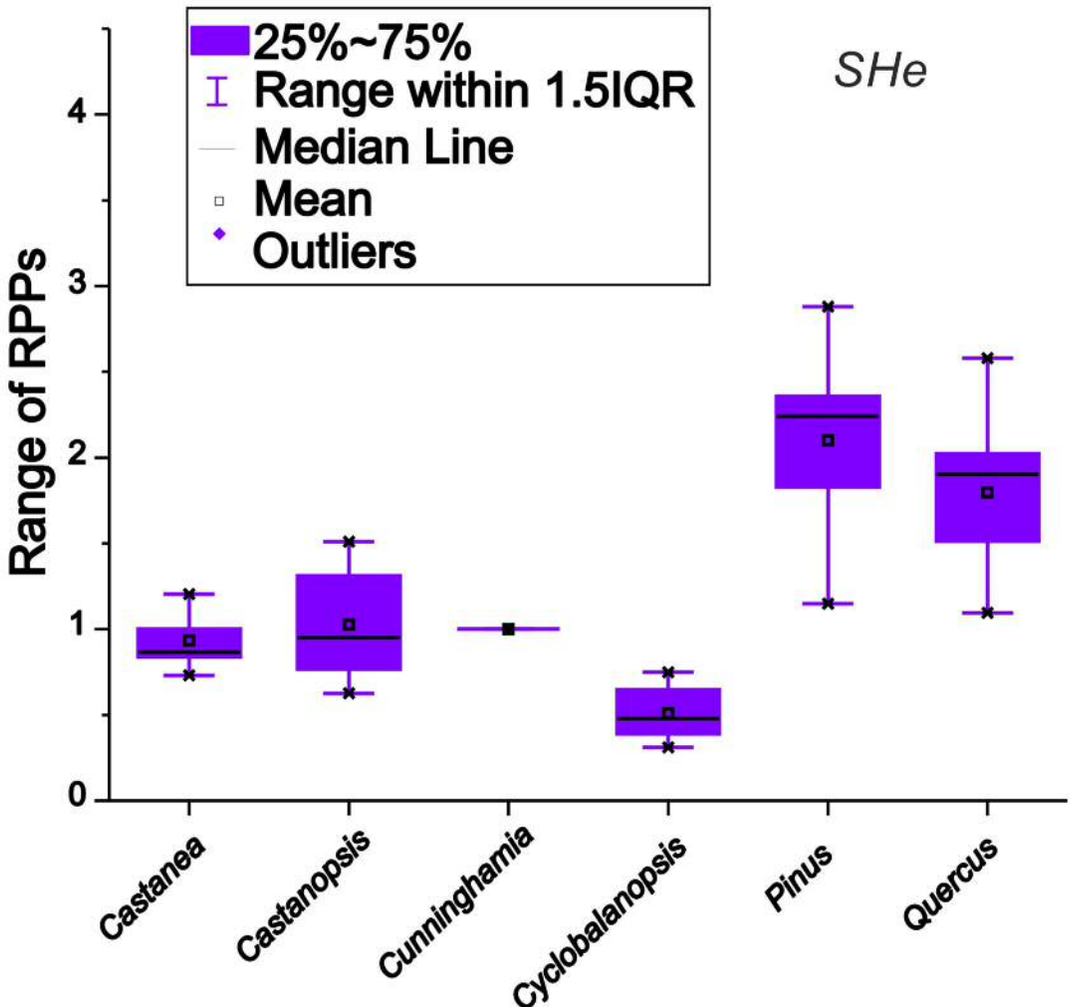
c



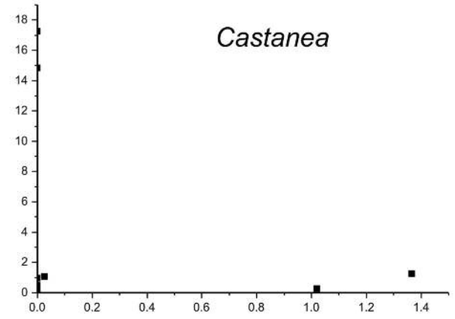
b

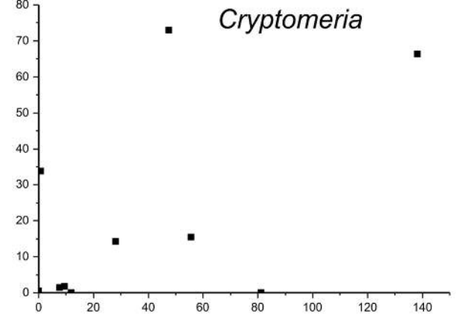


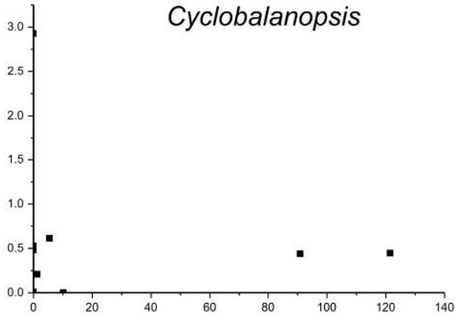
d

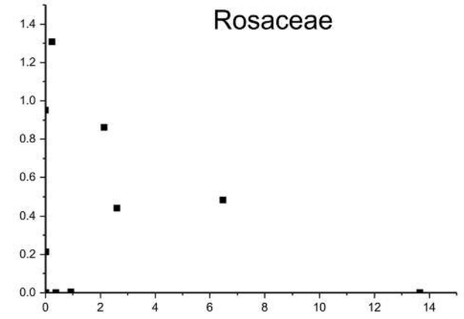


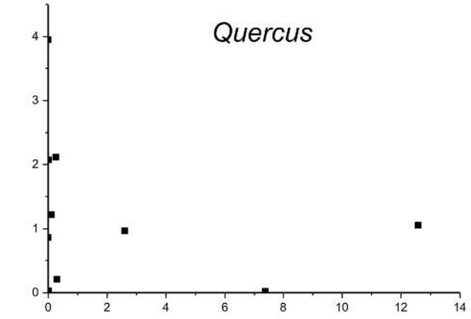
a

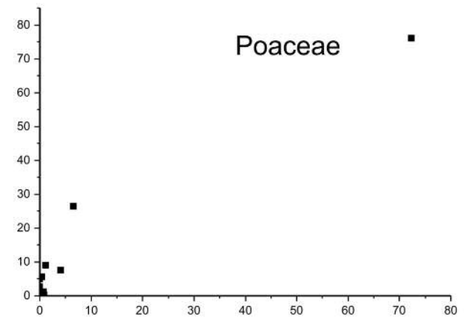

Castanea

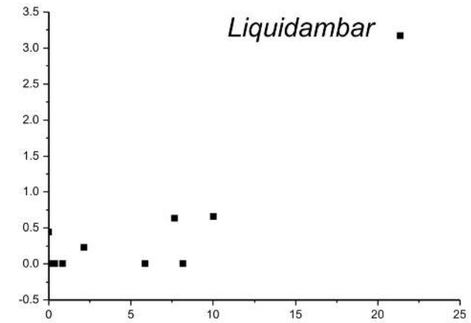

Cryptomeria

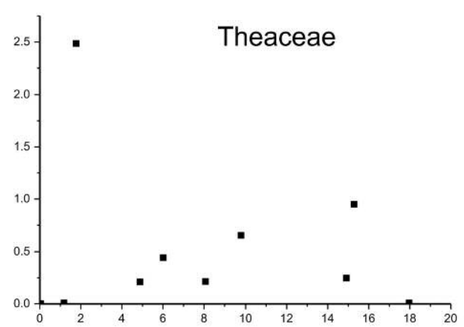

Cyclobalanopsis

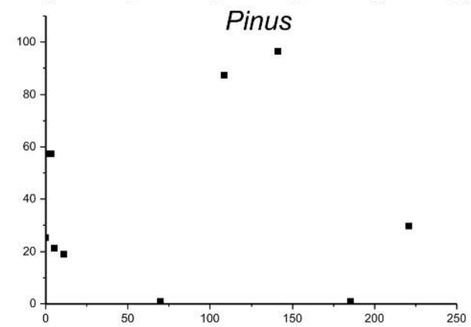

Rosaceae


Quercus

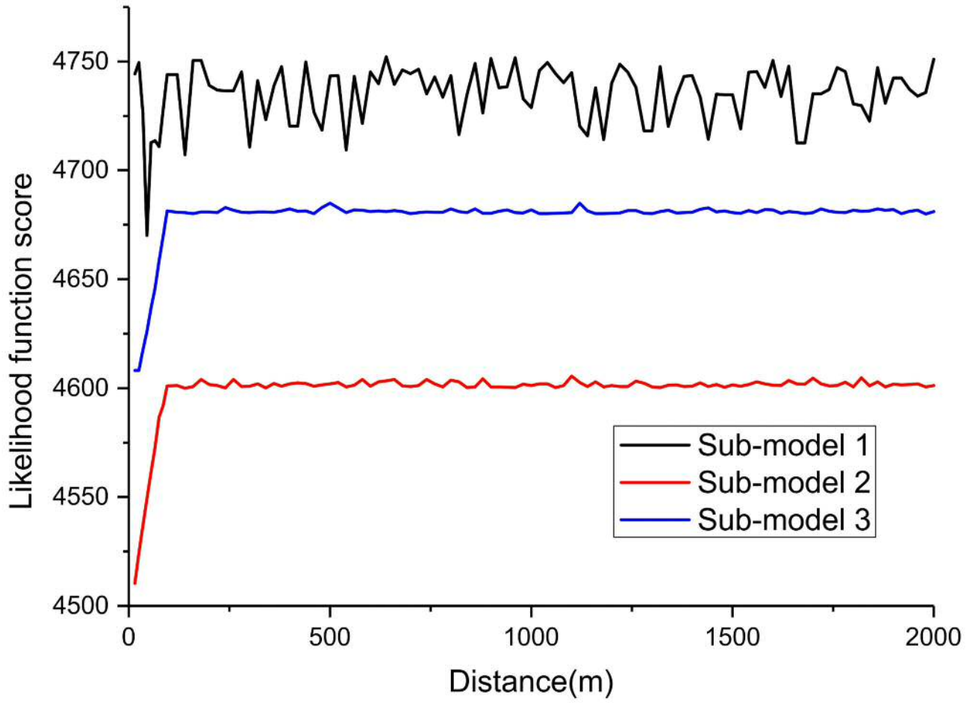

Poaceae


Liquidambar

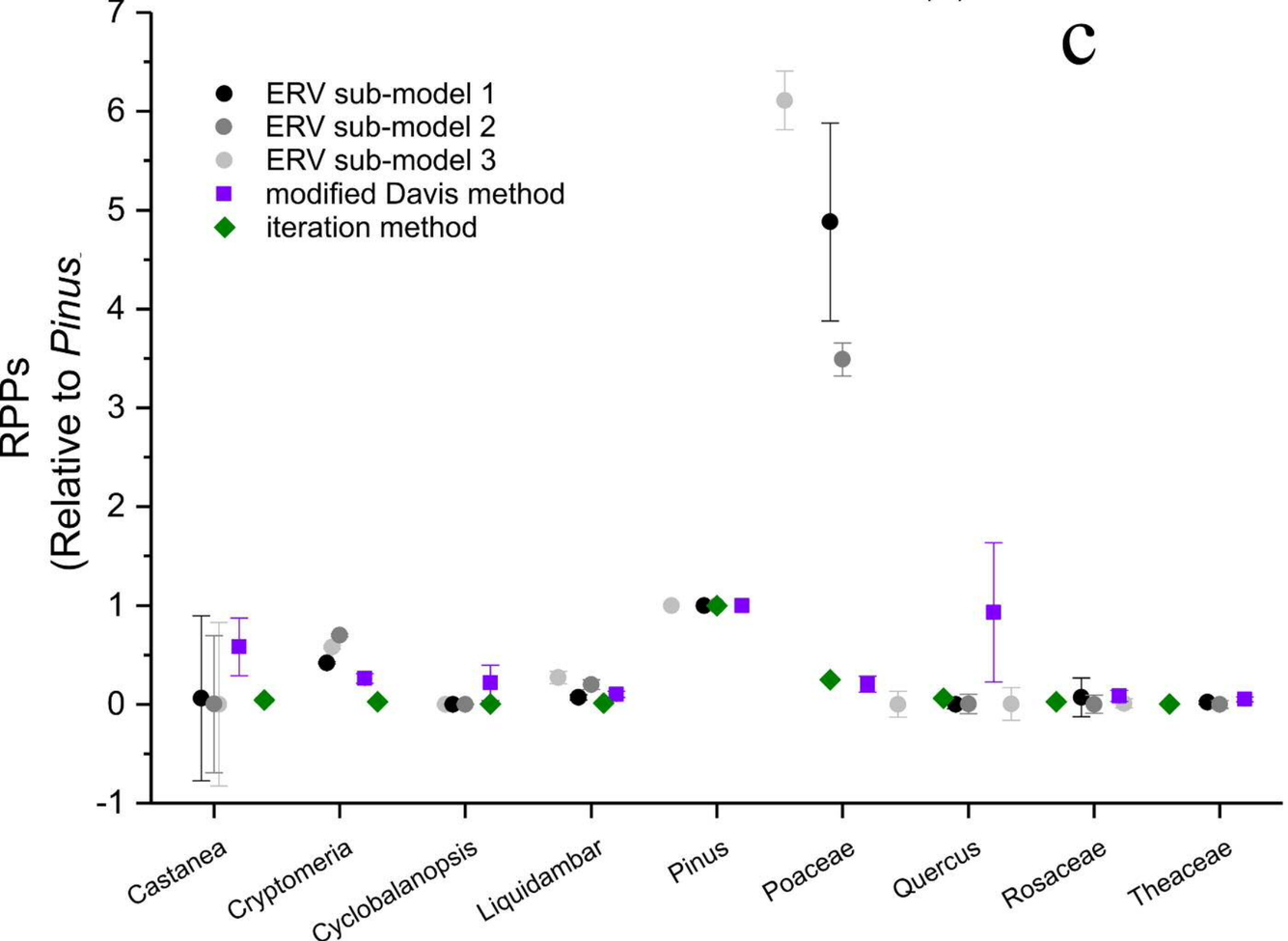

Theaceae


Pinus

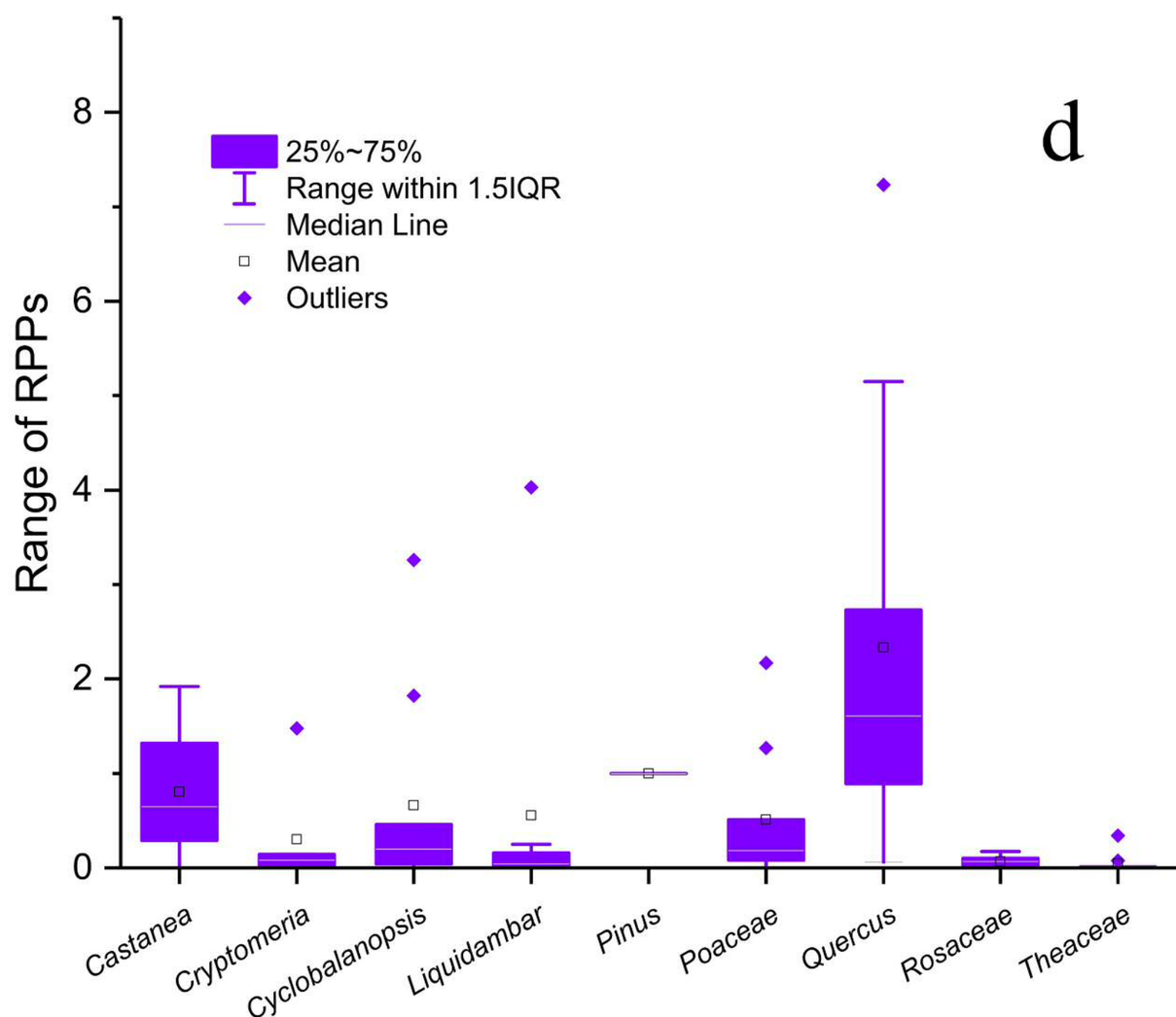
b


Likelihood function score
4750
4700
4650
4600
4550
4500
0
500
1000
1500
2000
Distance(m)
Sub-model 1
Sub-model 2
Sub-model 3

c


RPPs
(Relative to Pinus
ERV sub-model 1
ERV sub-model 2
ERV sub-model 3
modified Davis method
iteration method
Castanea
Cryptomeria
Cyclobalanopsis
Liquidambar
Pinus
Poaceae
Quercus
Rosaceae
Theaceae

d


Range of RPPs
25%~75%
Range within 1.5IQR
Median Line
Mean
Outliers
Castanea
Cryptomeria
Cyclobalanopsis
Liquidambar
Pinus
Poaceae
Quercus
Rosaceae
Theaceae

APPENDIX 1

The Modified Davis Method (MDM)

Davis (1963) defined the "R-value", a measure of pollen productivity, as the ratio between pollen proportion and vegetation abundance:

$$r_i = \frac{p_{ik}}{V_{ik}} \qquad \text{(Equation A1.1)}$$

She then argued that the ratio of the R-values for a pair of taxa, the Relative Pollen Productivity, should be constant between sites.

$$R_j i = \frac{r_i}{r_j} \qquad \text{(Equation A1.2)}$$

Davis (1963) measures vegetation as area of cover with no distance weighting applied, and pollen input from beyond the surveyed area (background pollen) is assumed to be negligible. We therefore modify Davis' approach by using distance weighted vegetation data collected to a distance many times larger than the likely RSAP, to include most of the possible sources of background pollen input. The calculation can be written as:

$$R_j i = \frac{p_{ik}}{v_{ik}} . \frac{v_{jk}}{p_{jk}} \qquad \text{(Equation 4)}$$

APPENDIX 2

Iteration Method (IM)

We begin with a dataset of k samples, each of which contain pollen percentage (p_{ik}) and dwpa values (v_{ik}) for m taxa, linked by equation 2.

Define n possible values of pollen productivity r_i (for example, in the simulation study we considered 6 possible values, 0.1, 0.5, 1, 2, 4 and 10), which give m^n possible RPP values for the full set of taxa.

We created a matrix of all possible combinations, then used each row to calculate estimated pollen loading values ($\hat{y}_{ik}$) from the v_{ik} values for the sample:

$$\hat{y}_{ik} = r_{in} \times x_{ik} \quad \text{(Equation A2.1)}$$

We converted these into estimated pollen proportions:

$$\hat{p}_{ik} = \frac{\hat{y}_{ik}}{\hat{y}_{\cdot k}} \quad \text{(Equation A2.2)}$$

Then compared estimated and actual pollen proportions for each of the k samples using Squared Chord Distance:

$$\text{SCD} = \sum_{i=0}^{m} \left(\sqrt{p_i} - \sqrt{\hat{p}_i}\right)^2 \quad \text{(Equation A2.3)}$$

And summed all k values to give a single Summed Squared Chord Distance measure of how well that particular set of values of r_i performed against the empirical data.

Values of SSCD were then compared for all possible m^n combinations of RPP values to identify the best combination of values. Where several combinations produced similar SSCD values, the estimates were averaged. Such combinations were identified by visual inspection of a plot of SSCD values in rank order.

R-code from simulation study

```
rm(list=ls())

gc()


test_x <- read.csv(file.choose())

head(test_x)


test_y <- read.csv(file.choose())

head(test_y)


library(tcltk)

pb <- tkProgressBar("progress","complete %", 0, 100)


startTime<-Sys.time()


nva=6

ntaxa=6

nsam=27


value<-c(0.1,0.5,1,2,4,10)

n=0

y_hat<-matrix(,nsam,ntaxa)

y_hat_pro<-matrix(,nsam,ntaxa)


maxRow = 19683

newFile = 1;
```

```
SSCDist<-matrix(,nrow=maxRow,ncol=ntaxa+1)
colnames(SSCDist)<-c("A","B","C","D","E","F","SSCD")
for (A in value){    ###the alpha of the first taxa changes according to value
 for (B in value){
  for (C in value){
   for (D in value){
    for (E in value){
     for (F in value){
                    alpha<-c(A,B,C,D,E,F)
                    y_hat<-t(alpha*t(test_x))
                    y_hat_pro<-y_hat/rowSums(y_hat)
                    distance<-sum((sqrt(y_hat_pro)-sqrt(test_y))^2)
                    SSCDist[n%%maxRow + 1,1:ntaxa]<-alpha
                    SSCDist[n%%maxRow + 1,ntaxa+1]<-distance
                    if (n%%maxRow == maxRow - 1 || n == nva^ntaxa - 1){
                            if (newFile == 1){
                                    newFile = 0
                                    write.table(SSCDist[1:(n%%maxRow + 1),],file="E:\\output-
SSCD.csv",sep=",",row.names=FALSE)
                            } else {
                                    write.table(SSCDist[1:(n%%maxRow + 1),],file="E:\\output-
SSCD.csv",append=TRUE,sep=",",row.names=FALSE,col.names=FALSE)
                            }
                    }

                    if (n %% round(nva^ntaxa / 100) == 0){
                            info<- sprintf("complete %d%%", round(n*100/nva^ntaxa))
```

```
                    setTkProgressBar(pb, n*100/nva^ntaxa, sprintf("progress (%s)", info),info)
                }
                n<-n+1
    }
   }
  }
 }
}
}

close(pb)
endTime<-Sys.time()
endTime-startTime

a<- read.csv(file.choose())   ##read the output-SSCD.csv file
b<-a[order(a[,7]),]          ##order the file according to the value of SSCD,in this study, the SSCD is in colum 7
head(b)
write.csv(b,file="E:\\order output-SSCD.csv")
```

APPENDIX 3

Land cover composition for simulation scenarios

Landscape	Grid	Colour	Community	Proportion in the grid	Pollen taxon	Proportion in each community	RPP	fallspeed
Homogeneous	Outer grid	Yellow	*Cunninghamia lanceolata* forest	6.1%	*Cunninghamia*	100%	2	0.016
		Green	*Pinus massoniana* forest	15.2%	*Pinus*	100%	4	0.063
		Blue-green	*Castanopsis sclerophylla* - *Cyclobalanopsis glauca* forest	54.6%	*Castanopsis*	60%	2	0.006
							1	0.012
					Cyclobalanopsis	40%		
		White	*Quercus fabri* forest	4%	*Quercus*	100%	4	0.016
		Pink	*Castanea seguinii* forest	20%	*Castanea*	100%	2	0.004
	Inner grid	Yellow	*Cunninghamia lanceolata* forest	6.3%	*Cunninghamia*	100%	2	0.016
		Green	*Pinus massoniana* forest	15.4%	*Pinus*	100%	4	0.063
		Blue-green	*Castanopsis sclerophylla* - *Cyclobalanopsis glauca* forest	54.3%	*Castanopsis*	60%	2	0.006
							1	0.012
					Cyclobalanopsis	40%		
		White	*Quercus fabri* forest	4%	*Quercus*	100%	4	0.016
		Pink	*Castanea seguinii* forest	20%	*Castanea*	100%	2	0.004
Nonhomogeneous	Outer grid	Yellow	*Cunninghamia lanceolata* forest	12.3%	*Cunninghamia*	100%	2	0.016
		Green	*Pinus massoniana* forest	43.4%	*Pinus*	100%	4	0.063
		Blue-green	*Castanopsis sclerophylla* - *Cyclobalanopsis*	32.5%	*Castanopsis*	60%	2	0.006
							1	0.012
					Cyclobalanopsis	40%		

		glauca forest					
	White	*Quercus fabri* forest	6.9%	*Quercus*	100%	4	0.016
	Pink	*Castanea seguinii* forest	4.9%	*Castanea*	100%	2	0.004
Inner grid	Yellow	*Cunninghamia lanceolata* forest	6.2%	*Cunninghamia*	100%	2	0.016
	Green	*Pinus massoniana* forest	14.7%	*Pinus*	100%	4	0.063
	Blue-green	*Castanopsis sclerophylla* - *Cyclobalanopsis glauca* forest	54.5%	*Castanopsis*	60%	2	0.006
				Cyclobalanopsis	40%	1	0.012
	White	*Quercus fabri* forest	4.3%	*Quercus*	100%	4	0.016
	Pink	*Castanea seguinii* forest	20%	*Castanea*	100%	2	0.004

APPENDIX 4

Vegetation communities within 100m of the empirical sampling points in the Meiling Mountains (sample locations shown in Figure 2; see text for details)

Sample Code	Vegetation types	Main species in the forest canopy	Main understory species
1	*Cryptomeria japonica* var. *sinensis* forest	*Cryptomeria japonica* var. *sinensis*	*Aster*, *Rhododendron simsii*, Theaceae spp.
2	*Pinus massoniana* - *Liquidambar formosana* - *Lithocarpus* sp. mixed forest	*Pinus massoniana*, *Liquidambar formosana*, *Lithocarpus* sp.	Ferns, moss *Buxus microphylla* subsp. *Sinica*, Theaceae spp.
3	*Pinus massoniana* - *Loropetalum chinense* forest	*Pinus massoniana*	*Rubus corchorifolius*, *Loropetalum chinense*, *Rubus parvifolius*, Theaceae spp.
4	Bamboo forest	*Phyllostachys edulis*	*Lindera aggregate*, Theaceae spp., *Adiantum capillus-veneris*, *Microlepia* sp.
5	Grassland	*Astragalus sinicus*, Poaceae spp.	*Astragalus sinicus*, Poaceae spp., *Oxalis corniculata*, *Erigeron* sp.
6	*Cunninghamia lanceolata* - *Pinus massoniana* - *Loropetalum chinense mixed* forest	*Cunninghamia lanceolata*, *Pinus massoniana*	*Loropetalum chinense*, Moss, *Adiantum capillus-veneris*, *Microlepia* spp., *Buxus microphylla* subsp. *Sinica*
7	*Cunninghamia lanceolata-Pinus massoniana-Loropetalum chinense* forest	*Cunninghamia lanceolata*, *Pinus massoniana*	*Loropetalum chinense*, Moss, *Dicranopteris pedata*, *Microlepia spp.*, Theaceae spp.
8	*Pinus massoniana* - *Loropetalum chinense* - *Quercus* spp. - *Castanea* sp. - *Cyclobalanopsis glauca* - *Platycarya*	*Pinus massoniana*, *Loropetalum chinense*, *Quercus* spp., *Castanea* sp.,	Moss, *Dicranopteris pedata*, *Microlepia* spp., *Carex* spp., Theaceae spp.

	strobilacea mixed forest	*Cyclobalanopsis glauca*, *Platycarya strobilacea*	
9	*Cyclobalanopsis glauca* - *Cunninghamia lanceolata* forest	*Cyclobalanopsis glauca*, *Cunninghamia lanceolata*	Moss, *Microlepia* spp., *Dicranopteris pedata*, *Buxus microphylla* subsp. *Sinica*
10	*Cyclobalanopsis glauca* - *Phyllostachys edulis forest*	*Cyclobalanopsis glauca, Phyllostachys edulis*	*Microlepia* spp., *Schima argentea*, moss

Some minor communities were not sampled, therefore are not listed in this table. They include small swamps which are mainly located at an elevation of 700-800m a.s.l., and support aquatic plants including *Juncus effusus*, *Sphagnum palustre*, *Ligularia japonica*, *Viola lactiflora*, *Polygonum thunbergii, Rotala rotundifolia* and *Salix chaenomeloides*, plantations (tea and rice) and settlements.

APPENDIX 5:

Fall speeds of the 9 selected main taxa from the Meiling Mountains

Pollen morphotype	Fall speed (m/s)	Published fall speed (m/s) and reference
Castanea	0.004	
Cryptomeria	0.015	
Cyclobalanopsis	0.012	
Liquidambar	0.034	
Pinus	0.063	0.031 (Eisenhut, 1961)
Quercus(deciduous)	0.016	0.035 (Eisenhut, 1961)
Rosaceae	0.009	
Theaceae	0.025	
Gramineae/Poaceae	0.030	0.035 (Sugita et al, 1999)