ORIGINAL ARTICLE

Biogeography of European land mammals shows environmentally distinct and spatially coherent clusters

H. Heikinheimo[1]*, M. Fortelius[2], J. Eronen[2] and H. Mannila[1,3]

[1]HIIT Basic Research Unit, Laboratory of Computer and Information Science, Helsinki University of Technology, PO Box 5400, Helsinki, FI-02015 HUT, Finland, [2]Department of Geology and Institute of Biotechnology, PO Box 64, University of Helsinki, Helsinki, FIN-00014, Finland, and [3]HIIT Basic Research Unit, Department of Computer Science, PO Box 68, University of Helsinki, Helsinki, FIN-00014, Finland

*Correspondence: H. Heikinheimo, Laboratory of Computer and Information Science, Helsinki University of Technology, PO Box 5400, FI-021014 HUT, Finland.
E-mail: hannes.heikinheimo@tkk.fi

ABSTRACT

Aim To produce a spatial clustering of Europe on the basis of species occurrence data for the land mammal fauna.

Location Europe defined by the following boundaries: 11°W, 32°E, 71°N, 35°N.

Methods Presence/absence records of mammal species collected by the Societas Europaea Mammalogica with a resolution of 50 × 50 km were used in the analysis. After pre-processing, the data provide information on 124 species in 2183 grid cells. The data were clustered using the *k*-means and probabilistic expectation maximization (EM) clustering algorithms. The resulting geographical pattern of clusters was compared against climate variables and against an environmental stratification of Europe based on climate, geomorphology and soil characteristics (EnS).

Results The mammalian presence/absence data divide naturally into clusters, which are highly connected spatially and most strongly determined by the small mammals with the highest grid cell incidence. The clusters reflect major physiographic and environmental features and differ significantly in the values of basic climate variables. The geographical pattern is a fair match for the EnS stratification and is robust between non-overlapping subsets of the data, such as trophic groups.

Main conclusions The pattern of clusters is regarded as reflecting the spatial expression of biologically distinct, metacommunity-like entities influenced by deterministic forces ultimately related to the physical environment. Small mammals give the most spatially coherent clusters of any subgroup, while large mammals show stronger relationships to climate variables. The spatial pattern is mainly due to small mammals with high grid cell incidence and is robust to noise from other subsets. The results support the use of spatially resolved environmental reconstructions based on fossil mammal data, especially when based on species with the highest incidence.

Keywords
Climate variables, clustering, Europe, mammalian fauna, metacommunity, presence/absence data, species distribution.

INTRODUCTION

Zoogeography, the subdivision of geographical space into regional subunits based on their fauna, has been a core activity of biology at least since the time of Wallace (1876). Originally its emphasis was on entire faunas and geographical units at the subcontinental or regional level (see the comprehensive review in Udvardy, 1969), but progress in the understanding of the spatial dynamics of metapopulations (Hanski, 1999; Hanski & Caggiotti, 2004) has shifted the focus towards single species and local scales. However, the question of how species are distributed in time and space remains a central research theme at all levels of scaling. Indeed, Holt & Keitt (2005) have recently argued that species borders qualify as a 'unifying theme' in ecology. After reviewing the current understanding of single-species borders, they make a plea for research that

doi:10.1111/j.1365-2699.2006.01664.x

considers borders from the point of view of entire metacommunities. While we appreciate that such work is usually focused on spatial dynamics and ecological relationships, we believe that the distribution patterns themselves also have considerable potential in this context. They are, after all, an outcome of just such dynamics and relationships, or the lack of them.

Whereas the spatial extents of plant metacommunities, especially in the sense of terrestrial vegetation, have been studied and modelled extensively (Hodgson *et al.*, 1999; Bonan *et al.*, 2002; Laurent *et al.*, 2004), the spatial distribution of animal communities has received relatively little attention since the heyday of quantitative zoogeography (e.g. Hagmeier & Stults, 1964; Hagmeier, 1966; Udvardy, 1969; Järvinen & Väisänen, 1973, 1980; Järvinen, 1979). Despite a wealth of research into the ecological factors involved in the control of the distributions of individual species or species pairs, there are few recent studies on the spatial distribution patterns within entire faunas. Even the discussion concerning niche-based vs. neutral explanations of community structure does not appear to have drawn on evidence of this kind (Chase, 2005), despite its obvious relevance for key assumptions about the relationship between species and their environment.

There is an additional reason to study the spatial properties of metacommunities: they provide a potential link to fossil-based studies at evolutionary time-scales, where the main data consist of presence or absence of taxa at specific locations during intervals measured in hundreds of thousands or millions of years. The metacommunity, regarded as the species pool from which local communities are drawn, is arguably (because of temporal and spatial averaging) the best match for the 'communities' or 'chronofaunas' (Olson, 1951) sampled in fossil material, and therefore connects patterns observed at evolutionary time-scales and those observed in present-day ecosystems. In particular, it is of interest to know whether spatial units at the subprovince level, typically resolvable in good-quality fossil data (Fortelius *et al.*, 2002), show evidence of coherence and environmental control in present-day data. Recent research in metacommunity structure and evolution (Leibold & Miller, 2004; Holyoak *et al.*, 2005) shows promise for producing data at this scale. It is therefore useful and timely for ecological as well as palaeontological purposes to investigate the spatial distribution characteristics of objectively defined clusters of present-day species.

Here we use distribution data for European land mammals to address the following questions: Do taxon aggregations at regional scales have meaningfully mappable spatial extent? If so, how robust are the patterns when using different clustering methods? How spatially coherent are the clusters? How do general properties of animals such as body size, trophic level, geographical range, abundance or conservation status influence the spatial patterns observed? Do the values of environmental variables differ between clusters, and does the geographical distribution of clusters correspond to independent, environmentally derived subdivisions of Europe? Here we show that the clusters found are coherent and robust, regardless of the methods used or the subsets analysed, and that they show strong relationships with environmental factors.

MATERIALS AND METHODS

Mammal data

We used mammal data (EMMA) collected by the Societas Europaea Mammalogica (http://www.european-mammals.org/) to prepare the *Atlas of European mammals* (Mitchell-Jones *et al.*, 1999). The data consist of presence/absence records of 194 mammal species for a set of 2670 grid cells covering Europe within boundaries 32°W, 35°E, 81°N, 30°N. The cell resolution of the grid is approximately 50 × 50 km, and the grid system is based on the Universal Transverse Mercator (UTM) projection and the Military Grid Reference System (MGRS). The grid system is the same as that used by the Atlas Florae Europaea (http://www.fmnh.helsinki.fi/english/botany/afe/index.htm).

For our purposes the data were pre-processed so that all records of bats, aquatic mammals, *Rattus* and *Mus*, and all mammals not native to Europe were excluded, except *Nyctereutes*, which has established a large population through recent natural dispersal. Excluding *Nyctereutes* had no significant effect on the results, and specifically did not change the special pattern for the areas covered by its range, such as Finland, the Baltic region or Bulgaria. *Mus* and *Rattus* were excluded because of their strong association with human habitation. The only effect of including these two genera was the spread of a coastal cluster uniting Ireland and the coast of Norway (see Results), to include other areas of low species diversity, especially the large Mediterranean islands, Greece, Transylvania and a scatter of cells in central Iberia. Pre-processing removal accounted for 70 of the 194 species in the data.

Sparse data for some cells present a problem in that the clustering methods tend to join cells with few species. This generates one spatially scattered cluster for which scarcity of species in the cells is the only common factor. Experimentation with cut-off-values from 0 to 20 species per cell suggested that removing cells with fewer than eight species eliminates this artefact without excluding too many grid cells from the analysis. With this setting, however, island areas, including Iceland, the Faeroe Islands, Svalbard and some small Atlantic islands, were excluded from the analysis.

For those cells with moderate or high number of species, changing the cut-off value from 0 to 20 did not affect the clustering results. This provides support for the robustness of the remaining clusters. Furthermore, this is in line with the fact that when comparing the clusterings with and without *Mus* and *Rattus* the biggest differences occur only in areas with few species. The final data set used in the study consisted of 2183 grid cells covering a window with boundaries 11°W, 32°E, 71°N, 35°N.

In the remaining data set, 'small mammals' are all species belonging to the orders Insectivora, Rodentia and

Lagomorpha. All others are regarded as 'large mammals', regardless of body size. This arbitrary grouping follows the common split among mammalogists into 'small mammal' vs. 'large mammal' specialists. This division is particularly relevant for comparison with fossil-based studies, where both the preservational regimes and the methodologies used for collection and study differ markedly between the groups, and where most of the literature is only concerned with one or the other.

All species were assigned to three trophic groups, carnivore, omnivore or herbivore, based on information collected from Nowak (1999) and occasionally other sources; assignments are available as supplementary material (see Supplementary Tables S1 & S2). The Societas Europaea Mammalogica retains the copyright of the data used. We use the term incidence to denote the percentage grid cell occurrence of a species, a concept close to 'commonness' as defined by Jernvall & Fortelius (2002). Incidence in this sense is primarily a measure of geographical range, but since range and local abundance are highly correlated (Hanski & Gyllenberg, 1997), incidence may be regarded as a reasonable proxy for abundance at scales that include most or all of the known range of a species. Incidence classes were defined by minimum coverage, thus 'present 10%' refers to species with a grid cell coverage of 10% or higher, and so on. We used data downloaded on 9 January 2006 from the IUCN Red List of Threatened Species (IUCN, 2004) to assign conservation status to species. We refer to the combined group of threatened and nearly threatened species as species at risk. The proportional overlap in species composition between the groups of the above-described groupings vary from no overlap to full overlap (Table 1).

Environmental data

Climate and elevation were obtained from Hijmans *et al.* (2005), available online at http://www.worldclim.org. The data consist of global climate layers in the 10′ ($18.6 \times 18.6 = 344$ km^2 at the equator) version of the data set. Climate and elevation values were associated to the UTM grid by taking an average of 10′ cells occurring within each UTM grid cell. We used three climate variables: average monthly mean temperature, average monthly precipitation and annual temperature range, that is, the difference between the maximum temperature of the warmest month and the minimum temperature of the coldest month. The records are from the period 1950–2000.

Table 1 Proportional overlap in species composition among subsets of data used to generate clusters

	s	l	h	o	c	r	nr	p10	p20	p30
a	0.73	0.27	0.41	0.27	0.32	0.18	0.82	0.41	0.35	0.23
s	1.00	0.00	0.43	0.27	0.30	0.17	0.83	0.34	0.28	0.19
l	0.00	1.00	0.35	0.26	0.38	0.21	0.79	0.59	0.53	0.35
h	0.76	0.24	1.00	0.00	0.00	0.20	0.80	0.33	0.24	0.18
o	0.73	0.27	0.00	1.00	0.00	0.15	0.85	0.42	0.39	0.27
c	0.68	0.33	0.00	0.00	1.00	0.17	0.82	0.50	0.45	0.28
r	0.68	0.32	0.45	0.23	0.32	1.00	0.00	0.27	0.23	0.09
nr	0.74	0.26	0.40	0.27	0.32	0.00	1.00	0.44	0.37	0.26
p10	0.61	0.39	0.33	0.27	0.39	0.12	0.88	1.00	0.84	0.57
p20	0.58	0.42	0.28	0.30	0.42	0.12	0.88	1.00	1.00	0.67
p30	0.59	0.41	0.31	0.31	0.38	0.07	0.93	1.00	1.00	1.00

a, all species; s, small mammals, l, large mammals; h, herbivora; o, omnivora; c, carnivora; r, at risk; nr, not at risk; p10, present 10%; p20, present 20%; p30, present 30%.

Clustering methods

We used the well-known clustering methods, *k*-means (also known as ISODATA) (Duda *et al.*, 2000; Theodoridis & Koutroumbas, 2003) and probabilistic Bernoulli mixture modelling using the expectation maximization (EM) algorithm (Everitt & Hand, 1981; Cadez *et al.*, 2000; McLachlan & Peel, 2000; Hand *et al.*, 2001), to obtain a clustering of the grid cells.

The *k*-means clustering method is based on a simple iterative process. The method is initialized using a random assignment of cluster centres. The first step of the procedure is to take each point in the data set and associate it with the nearest cluster centre in terms of Euclidean distance. The second step is to recalculate each cluster centre by assigning it to the mean of the data points associated with it. By repeating these two stages the cluster centres change their location step by step and after a sufficient number of iterations converge to a locally optimal position in the data space. The final clustering is obtained by associating each data point with the nearest converged cluster centre. The EM clustering algorithm for Bernoulli mixture modelling works in a similar way but extends the approach to a probabilistic framework. The crucial difference from the *k*-means clustering is thus in the way in which the data points are assigned to the clusters. More precisely, the EM algorithm computes probabilities of cluster memberships for each data point and cluster centre. Then each cluster centre is recomputed as an average of the points, weighted by the probability of cluster membership for this cluster. Again the two stages are repeated until convergence. Here the final clustering is obtained by associating each data point with the most probable cluster. The clustering methods use only the species presence/absence data as input.

The similarity of the clusterings generated by the two methods employed was compared using the kappa statistic (Monserud & Leemans, 1992). To evaluate the kappa statistic we used the qualitative guidelines of Monserud & Leemans: kappa less than 0.2 represents very poor agreement, 0.2–0.4 is poor agreement, 0.4–0.55 is fair agreement, 0.55–0.7 is good agreement, 0.7–0.85 is very good agreement and a value over 0.85 is excellent agreement. As a technical detail, note that before the kappa statistic can be computed, it must be decided which clusters correspond to each other in the two clusterings. This matching was done so that the aggregate geographical overlap between the matched clusters was maximized. For this we used the minimum-cost perfect matching algorithm described in detail in Kleinberg & Tardos (2005).

The choice of the number of clusters was based on the climatic stratification (EnS) presented in Metzger *et al.* (2005), who divide the geographical region of their study into 13 environmental zones (EnZ). However, the areas covered by their Anatolian cluster are not included in our data, leaving 12 environmental zones. The choice of cluster number in our analysis was thus 12.

Environmental comparisons of clusters

We used one-way ANOVA (Krzanowski, 1988; Rencher, 2002) to test whether the generated clusters differ significantly in the values of environmental variables. For each clustering a set of ANOVAS were conducted separately for each of the four environmental variables with respect to each pair of clusters.

Spatial analysis of clusters

To quantify the spatial coherence in the clusters we defined a neighbour relation between the grid cells. For each cell we defined eight spatial neighbours as the eight bordering cells. We concluded that a cell fulfils an eight-neighbour coherence condition if all of its eight spatial neighbours belong to the same cluster as the cell itself. We measured the spatial coherence of the clusters with the number of cells fulfilling this condition. Because coastal and insular cells have less than eight neighbours, the condition is considered fulfilled if all of the existing neighbours belong to the same cluster. To evaluate the robustness of the spatial pattern with respect to the number of clusters, we generated a sequence of clusterings with different numbers of clusters from 2 to 13 for the all-species data set.

RESULTS

Comparison of clusterings

The kappa values computed between the least error clusterings of the *k*-means and the probabilistic modelling method show that the two clustering methods produce similar results (Table 2). Using the qualitative guidelines of Monserud & Leemans (1992) for kappa comparison the species data sets 'all', 'large', 'omnivora', 'carnivora', 'present 10%', 'present 20%' and 'present 30%' have very good agreement. All other sets have good agreement. The species set for which the methods disagree the most is the 'at risk' set (threatened or near-threatened in the IUCN classification).

There is fair agreement with an environmental stratification of Europe based on non-biotic information (EnS) (Table 2). The species sets 'all' and 'herbivora' give a kappa value of 0.4 when comparing the EnS environmental zones with the *k*-means clustering of the smallest error. A similar comparison undertaken using the probabilistic clustering of the smallest error gave fair agreements for the sets 'small mammals', 'large mammals', and 'herbivora', with herbivores yielding the overall best kappa of 0.47. For both *k*-means and probabilistic modelling the other species sets give kappa values of less than 0.4 with again the set 'at risk' having the lowest, around 0.24.

Table 2 Stability of the clustering results and strength of spatial agreement between the EnS stratification (Metzger *et al.*, 2005). The first column shows the kappa statistic between the clustering results of the *k*-means and probabilistic EM clustering method. The second and the third columns show the kappa statistic between the EnS stratification and the *k*-means and the probabilistic EM clusterings. The kappa values have been computed using the clusterings with the smallest error out of 100 runs for each species set

Species set	Kappa *k*-means vs. EM	Kappa *k*-means vs. EnS	Kappa EM vs. EnS
All species	0.76	0.40	0.37
Small species	0.69	0.37	0.43
Large species	0.72	0.39	0.40
Herbivora	0.60	0.40	0.47
Omnivora	0.72	0.29	0.33
Carnivora	0.79	0.35	0.37
At risk	0.52	0.24	0.26
Not at risk	0.66	0.35	0.37
Present 10%	0.79	0.35	0.36
Present 20%	0.81	0.38	0.37
Present 30%	0.70	0.35	0.32

Mammal cluster results

The clusters are spatially very coherent (well-connected), even though the clustering methods use only the presence/absence data (Fig. 1). The most coherent species sets are 'all species', 'species not at risk' and 'small mammals'. Depending on the clustering method these sets have about 1148–1300 cells fulfilling the eight-neighbour coherence condition. The clusterings for the sets 'present 10%' and 'present 20%', 'herbivora' and 'carnivora' are only slightly less coherent, with 928–1147 cells fulfilling the same condition. Likewise, the sets 'present 30%', 'omnivora' and 'large mammals' have 735–1141 cells fulfilling the coherence condition. The set 'at risk' is clearly less coherent, with only 359 cells with *k*-means and 610 cells with probabilistic modelling fulfilling the coherence condition.

The spatial coherence of clusters is not well explained by the number of species in the data set (Fig. 1a): the data sets 'all species', 'species not at risk', 'small mammals' and 'present 20%' all vary in the number of species and yet have high spatial coherence. Also, the species sets 'present 10–30%', 'herbivora', 'carnivora', 'omnivora' and 'large mammals' vary in spatial coherence although they contain similar numbers of species. Furthermore, there is no relationship between spatial coherence and the incidence of species in the sets (Fig. 1b).

Increasing cluster number sequentially from 2 to 13 showed that as the number of clusters increases, their spatial expressions tend to successively divide into smaller units within the pattern already established (Fig. 2). The initial split into two clusters separates the continental margins from a large central

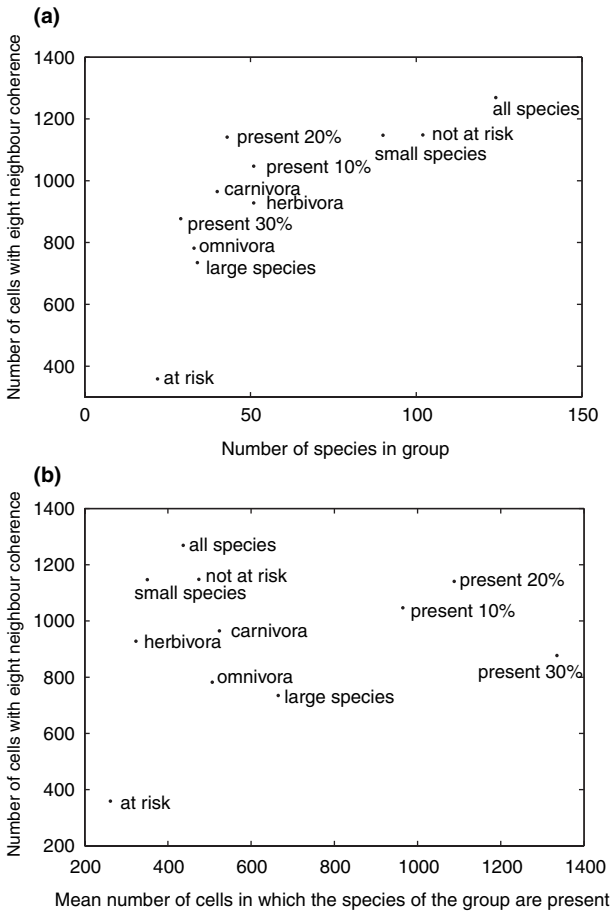

Figure 1 Spatial coherence of the *k*-means clusterings compared to the average incidence and size of the studied species subsets. Spatial coherence of a clustering is measured as the number of cells fulfilling the eight-neighbour coherence condition. The coherence condition is fulfilled for a cell if all eight of its spatial neighbours belong to the same cluster as the cell itself. (a) Number of species vs. the number of cells fulfilling the eight-neighbour coherence condition in each of the studied species subset. (b) Mean number of cells in which the species of the group are present (incidence) vs. the number of cells fulfilling the eight-neighbour coherence condition in each of the studied species subset.

area along boundaries that persist roughly throughout the sequence of splitting. The second split (at three clusters) divided the peripheral cluster north–south. The next split (at four clusters) divided the centre, followed at five clusters by a north–south split of Scandinavia. At six clusters the southern-most part is split east–west, at seven the eastern-most part is split north–south. At eight clusters a unit connecting the southern half of Finland and Estonia appears, at nine clusters Ireland joins a unit that henceforth persistently extends north along the Norwegian coast. From this stage onwards, the pattern of northern Europe did not change, the remaining steps successively creating distinct clusters for the Alps, southern France and finally the Italian Peninsula.

The edges of the spatial distribution of 12 clusters (Figs 3 & 4) are reasonably constant, with boundaries usually following major environmental features. Areas separated by major bodies of water tend to be separated, but terrestrial topographic and climatic features are also evident. The Alps are identified as a separate spatial unit from 11 clusters onward in the 'all species' data set (Fig. 2) and in 6 of the 11 clusterings shown in Fig. 4, whereas the boundary between the Iberian Peninsula and the rest of Europe follows the southern boundary of the Pyrenees and the Cantabrian Mountains. Similarly, the very stable boundaries that cross Finland at about 65°N and Scandinavia at about 60°N (the Limes Norrlandicus of Linnaeus) are biogeographically well known (Udvardy, 1969; Järvinen & Väisänen, 1973). The subdivisions of continental Europe are somewhat more variable, but several cluster boundaries nonetheless occur frequently, notably the nearly straight, east–west line that crosses Germany and Poland around 52–54°N. This stable feature, seen from seven clusters onwards (Fig. 2), does not correspond with any previously recognized boundary, but it does approximate the boundary between the low-lying coastal area and the more elevated regions further inland. Several boundaries are also close to hybridization zones based on genetic evidence (Hewitt, 2000). In addition to the Alps, the Pyrenees and the Limes Norrlandicus, these zones correspond to the boundary that runs across France from Languedoc to Brittany and the boundary between Estonia and the more southern Baltic states.

By and large, the most constant boundaries are in an east–west direction, suggesting that temperature is a primary underlying factor. The least coherent region is in south-east Europe, roughly corresponding to Bulgaria, Romania and the former Yugoslavia (Fig. 4). This probably reflects the fact that species in this part of Europe have more fragmented ranges. Whether this reflects poorer data quality or the fact that the area lies near the boundaries of several traditionally recognized biogeographical regions (e.g. Udvardy, 1969, Figs 5–11) cannot be determined from the available data, but the incompleteness of the data for Romania and Bulgaria is a known issue in the EMMA data (A. J. Mitchell-Jones, pers. comm.).

A comparison between 'small' and 'large' mammals shows that, while many of the same spatial features are seen for both data sets, large mammals yield much less coherent clusters (Figs 1 & 4). Indeed, the cluster coherence seen for large mammals is the lowest seen in any subset save that of species at risk. The pattern for small mammals, in contrast, is similar to those seen for all species and all species except those at risk, in spatial pattern as well as coherence (Table 3, Fig. 1). Clusters for small mammals also show physical geographical features, such as mountain chains and water bodies, more clearly (Fig. 4).

The spatial pattern of the clusters for the trophic groups 'carnivora', 'herbivora' and 'omnivora', have a strong overall similarity with each other and with the pattern of all species (Table 3, Fig. 4). The herbivore pattern corresponds more closely with the environmentally based EnS stratification (Metzger *et al.*, 2005) than does any other subgroup (Table 2),

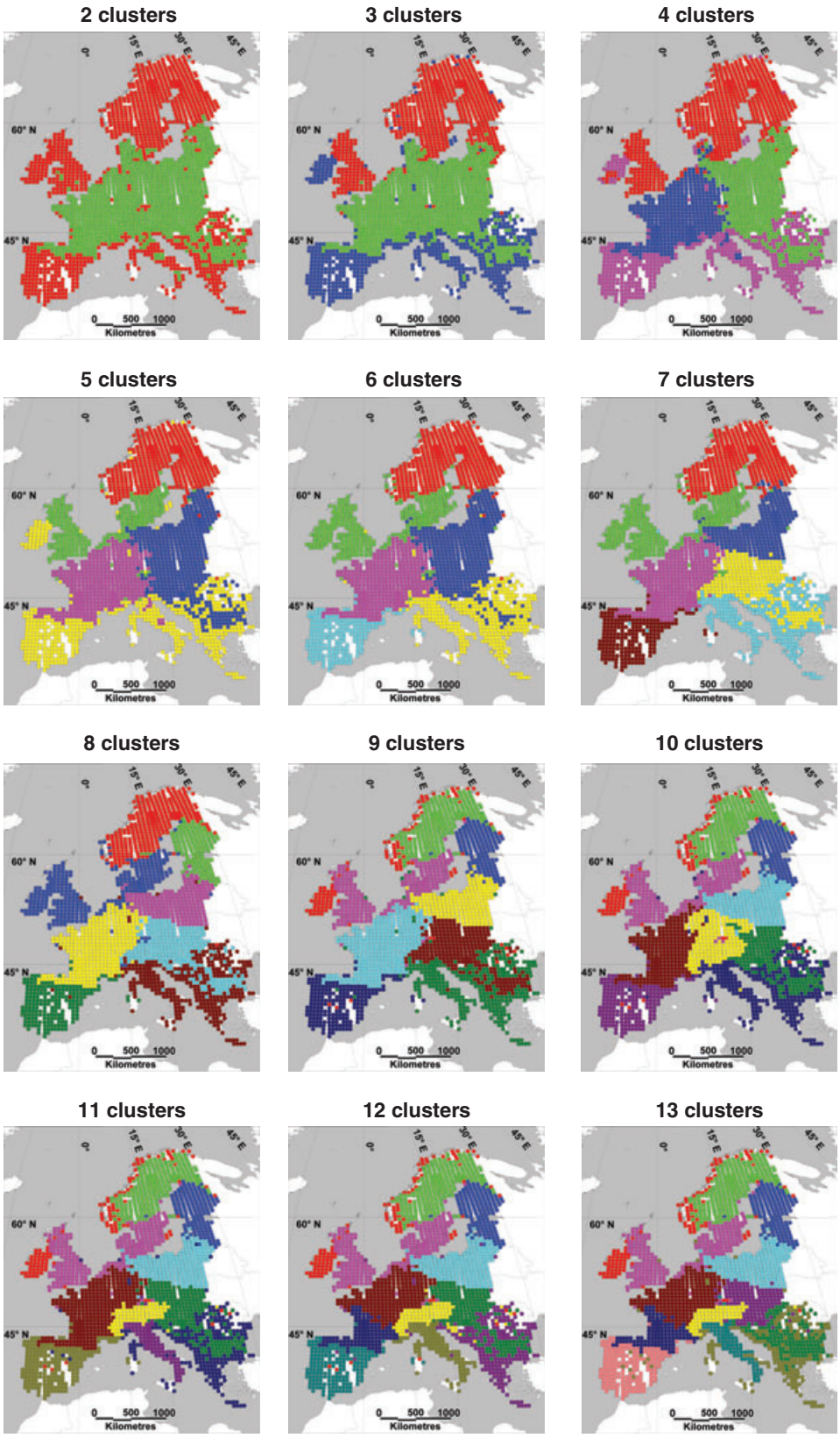

Figure 2 The sequence of clusterings of the mammal data cells with clusters from 2 to 13 computed with the 'all species' set. The figure shows the *k*-means clustering with the smallest error out of 100 runs for each cluster number assignment. The maps are plotted using the Mollweide (equal-area) NAD27 projection. The colours are used only to distinguish the clusters within each image and do not imply a one-to-one matching of clusters between images.

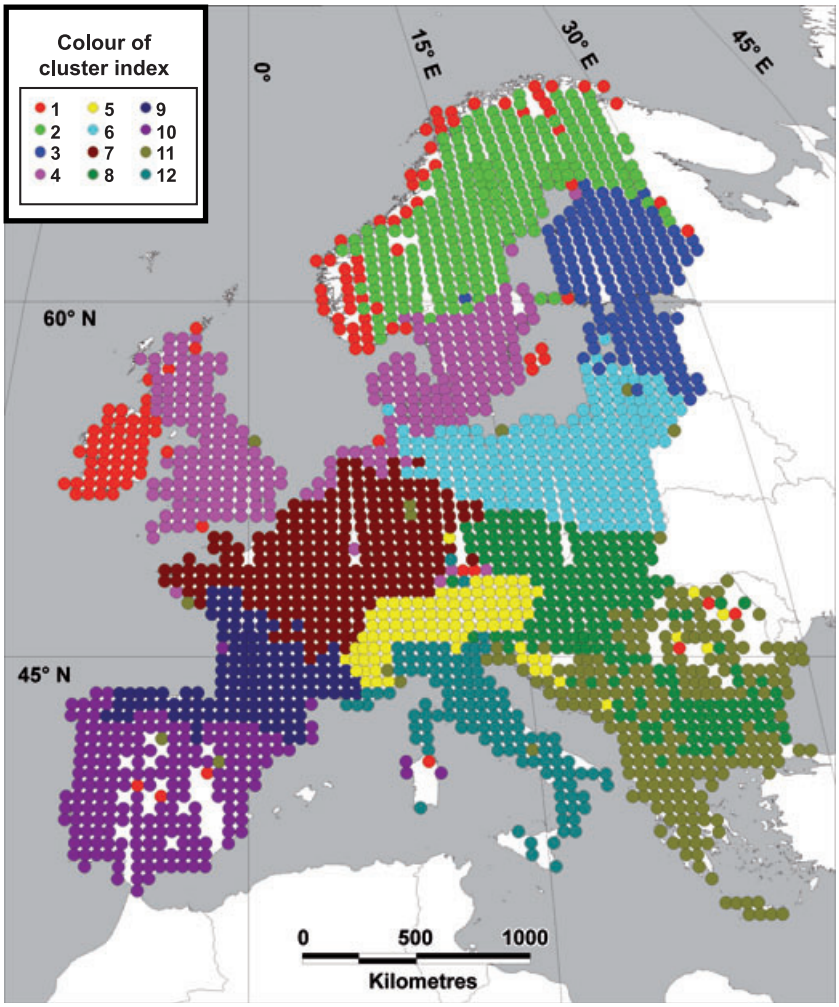

Figure 3 The *k*-means clustering of the mammal data cells in 12 clusters with the 'all species' set. The clustering is the best out of 100 clustering runs in terms of squared error. The cells are projected on to the map with the Mollweide (equal-area) NAD27 projection.

possibly because herbivore distributions are most directly influenced by the maritime–continental climate gradient.

The species with the highest grid cell incidence give more coherent clusters than other groups (Fig. 1). Those with an incidence of 10–20% give coherence values approaching those of all species and small mammals, but higher incidence values give lower coherence, perhaps because the species with the highest incidence are few and widespread. The subset of species 'at risk' gives spatially the least coherent clusters found in this study, even less coherent than seen for large mammals (Fig. 1).

The regional divisions identified by the clusterings show significant differences in the values of basic climate variables and elevation (Table 4). All cluster pairs in the 'all species' clustering seen in Fig. 3 differ significantly in at least two environmental variables, and most cluster pairs differ in all of the variables (Table 4a,b). For almost all groupings temperature is the variable for which the cluster pairs have the most significant differences (Table 4c). For precipitation, the number of significant differences is also high. For all environmental variables the set 'species at risk' has the smallest number of significantly different cluster pairs, while the species set with the largest number of significant differences is different for each considered variable. However, more important than these relatively minor differences is the high overall percentage of significant differences. The results of the ANOVA tests complete with *P*-values for all of the species groupings are provided as Table S3 in the supplementary material.

DISCUSSION

We find that Europe can be divided into coherent subregions based on the distributions of mammal species. We also find a high degree of geographical coherence displayed by the clusters, and consistency in the basic spatial pattern among non-overlapping subsets of the data and despite changes in the number of clusters. These observations, in combination with the environmental contrast observed between the clusters and the concordance of the geographical cluster pattern with the EnS environmental stratification strongly suggest that the clusters represent real biological units rather than arbitrary constructs generated by the clustering algorithms. We take this to indicate that, even in present-day Europe with its long history of intensive human presence, the main controls on mammalian metacommunity distributions remain

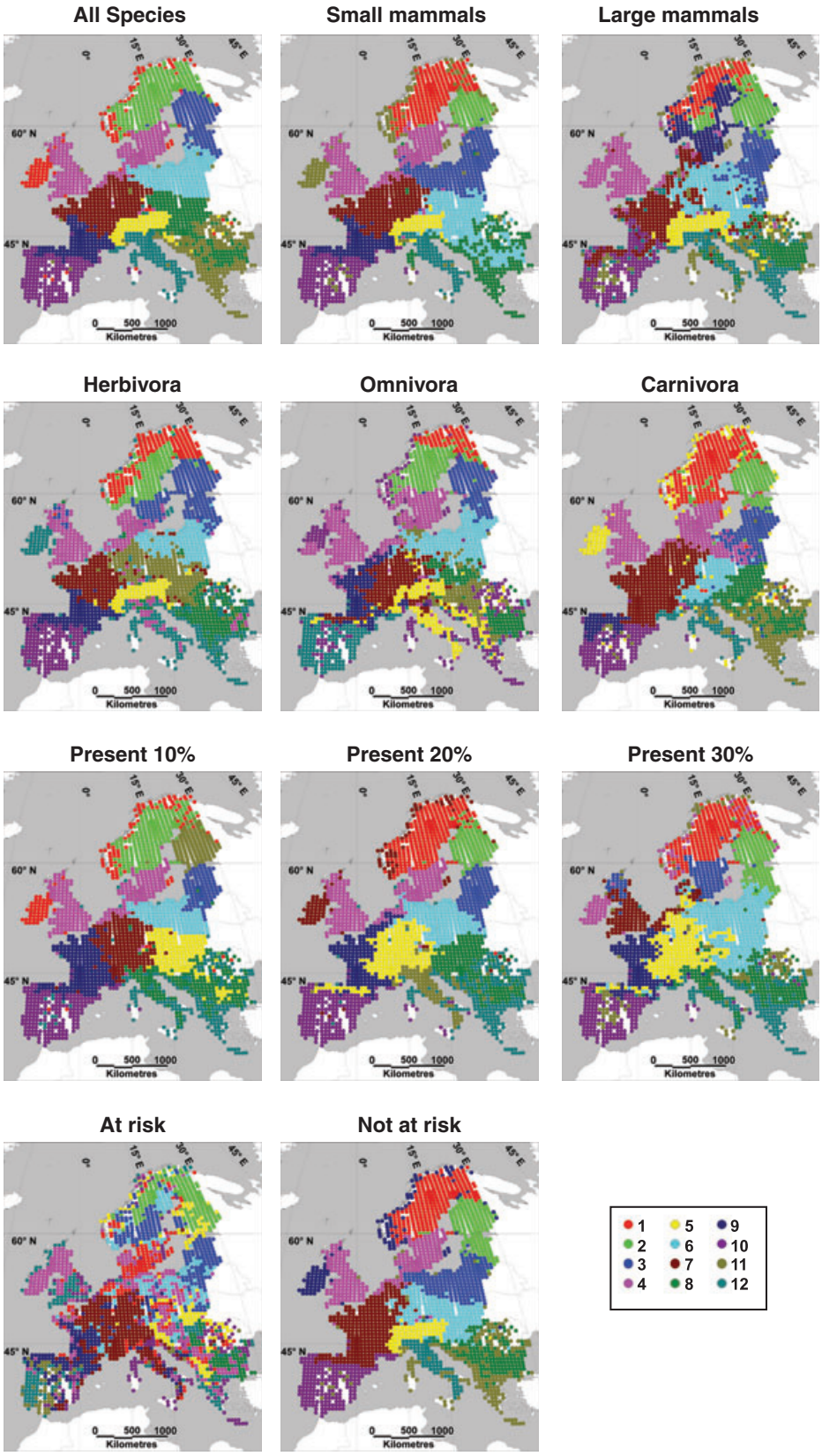

Figure 4 The *k*-means clusterings of the mammal data cells into 12 clusters with respect to the species sets: all species, small mammals, large mammals, herbivora, omnivora, carnivora, at risk, not at risk, present 10%, present 20% and present 30%. The clusterings are the best out of 100 clustering runs in terms of squared error. Presentation as in Fig. 2.

Table 3 Strength of spatial agreement between the *k*-means clusterings of the mammal data cells with different species sets, measured with the kappa statistic. The values have been computed using the *k*-means clusterings with the smallest error out of 100 runs for each species set. The number of clusters is 12

	s	l	h	o	c	r	nr	p10	p20	p30
a	0.87	0.44	0.60	0.59	0.72	0.29	0.78	0.71	0.74	0.58
s	–	0.43	0.56	0.57	0.68	0.27	0.73	0.69	0.71	0.56
l	–	–	0.47	0.42	0.43	0.28	0.53	0.45	0.48	0.49
h	–	–	–	0.50	0.45	0.28	0.63	0.55	0.46	0.45
o	–	–	–	–	0.45	0.25	0.58	0.58	0.59	0.51
c	–	–	–	–	–	0.25	0.66	0.59	0.66	0.50
r	–	–	–	–	–	–	0.27	0.32	0.29	0.27
nr	–	–	–	–	–	–	–	0.74	0.67	0.58
p10	–	–	–	–	–	–	–	–	0.81	0.62
p20	–	–	–	–	–	–	–	–	–	0.62
p30	–	–	–	–	–	–	–	–	–	–

a, all species; s, small mammals, l, large mammals; h, herbivora; o, omnivora; c, carnivora; r, at risk; nr, not at risk; p10, present 10%; p20, present 20%; p30, present 30%.

predominantly related to natural factors such as topography and climate. There is limited evidence that the controls differ in their effects on different subsets of species, but the overall similarity between patterns derived from non-overlapping subsets, including different trophic groups, suggests that direct influence from the physical environment is the most important control for all mammals except, perhaps, those currently at risk. This is further supported by the fact that even when the clusters become spatially incoherent, the climatic differences between them remain.

From metapopulation theory (Hanski, 1999; Hanski & Caggiotti, 2004) it might be expected that the species with the highest regional incidence largely correspond to the core species of local populations, and thus that a metacommunity made up of core species should be a more stable and distinctly bounded unit than a metacommunity that includes rare (satellite) species. The clustering behaviour described in this paper is consistent with this expectation: the most coherent clustering is actually observed for the full set of species, but selecting only species with high incidence produces coherent clusters, as discussed above. It is perhaps especially significant that noise from the species with low incidence does not

Table 4(b) Results of the ANOVA tests for each pair of clusters shown in Fig. 3, separately for each climate variable and for elevation. Variables that are significantly different (P-value < 0.05) in pairwise comparisons of clusters: a, all climate variables and elevation; e, elevation; p, precipitation; t, temperature; r, annual temperature range

	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12
C1	a	a	e,p,t	e,r	a	p,t,r	a	a	a	a	p,t,r
C2	–	e,p,t	a	a	a	a	p,t,r	p,t,r	e,t,r	p,t,r	a
C3	–	–	p,t,r	a	e,t,r	a	a	a	e,t,r	a	a
C4	–	–	–	a	p,t,r	e,t,r	a	a	a	a	e,t,r
C5	–	–	–	–	a	a	a	a	e,p,t	a	a
C6	–	–	–	–	–	a	a	a	e,t,r	e,p,t	a
C7	–	–	–	–	–	–	a	e,p,t	a	a	e,t,r
C8	–	–	–	–	–	–	–	p,t,r	a	t,r	a
C9	–	–	–	–	–	–	–	–	p,t,r	p,r	a
C10	–	–	–	–	–	–	–	–	–	a	a
C11	–	–	–	–	–	–	–	–	–	–	t,r

Table 4(c) The percentage of cluster pairs that differ significantly for each environmental variable in each of the studied *k*-means clusterings

	Elevation	Precipitation	Temperature	Temperature, annual range
All species	0.8	0.85	0.97	0.92
Small mammals	0.82	0.92	0.95	0.91
Large mammals	0.86	0.76	0.97	0.89
Herbivora	0.83	0.79	0.97	0.88
Omnivora	0.86	0.82	0.94	0.92
Carnivora	0.83	0.86	0.92	0.92
At risk	0.64	0.74	0.89	0.79
Not at risk	0.80	0.86	0.98	0.88
Present 10%	0.80	0.89	0.94	0.95
Present 20%	0.83	0.91	0.92	0.94
Present 30%	0.83	0.74	0.94	0.88

Table 4(a) The mean values and standard deviation (SD) of climate variables and elevation in each of the clusters in Fig. 3. C1, cluster indexed with colour number 1 in Fig. 3; C2, cluster indexed with colour number 2 in Fig. 3, and so on

	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12
Elevation(m)												
Mean	281	409	92	99	1167	105	249	415	454	524	392	332
SD	321	273	56	103	542	65	197	304	441	354	313	249
Precipitation (mm year^{-1})												
Mean	1135	666	611	805	1148	607	782	704	874	639	734	768
SD	458	214	45	275	230	67	126	165	144	242	243	137
Mean temperature (°C)												
Mean	5.93	0.70	3.70	7.94	5.98	7.55	9.57	9.11	11.32	14.12	11.43	13.54
SD	4.10	2.14	1.52	1.50	2.92	0.95	1.14	1.84	1.95	2.33	3.06	1.89
Temperature annual range (°C)												
Mean	21.2	32.2	32.7	22.0	26.6	28.9	23.9	29.1	24.1	26.2	28.7	25.1
SD	6.27	4.74	2.29	3.94	2.42	2.46	2.53	1.64	2.29	4.14	4.67	2.67

decrease the coherence of clusters, which are structured by the species with the highest incidence, especially of small mammals. We therefore propose that the clustering patterns reflect the spatial distribution of meaningfully distinct mammalian metacommunities.

Our results thus support the interpretation that spatially and structurally distinct assemblages seen in the mammalian fossil record (Bernor *et al.*, 1979, 1996; Bernor, 1983, 1984) reflect ecologically distinct biogeographical units (e.g. palaeobiomes), bounded by environmental conditions. Except for data severely altered by the various processes that transform a living community into a fossil assemblage (Badgley *et al.*, 1995) or by sampling biases (Alroy *et al.*, 2001), it is thus reasonable to conclude that spatially resolved palaeoenvironmental reconstructions based on fossil mammals (Barnosky & Carrasco, 2002; Fortelius *et al.*, 2002, 2004; Barnosky *et al.*, 2003) reflect underlying relationships between the once-living mammalian metacommunities and their environments (see also Damuth, 1982). The fact that a fossil collection virtually always represents more time and a greater geographical area than the living community sampled from a particular spot may, in this context, be more of an advantage than a problem, as it tends to reduce the noise from small-scale variation (Jernvall & Fortelius, 2004). Our results also imply that small mammals are most useful for regional subdivisions, whereas large mammals may be more useful for stratigraphic correlation over larger areas, and herbivores may be the most useful group for reconstructing environmental conditions.

The fact that the species with the highest incidence give a pattern almost identical to that based on all species is particularly encouraging for connecting fossil-based studies with data from living ecosystems. Whether because they are locally abundant or widespread, or in most cases probably for both reasons (Hanski & Gyllenberg, 1997), the most 'common' species are the ones predominantly sampled in the fossil record and the ones that most usefully describe evolutionary trends (Jernvall & Fortelius, 2002; Vermeij & Herbert, 2004). Finally, our findings are in line with those reported by McGill *et al.* (2005), suggesting that deterministic forces are at work in shaping mammalian communities (and therefore metacommunities) on evolutionary time-scales.

ACKNOWLEDGEMENTS

We thank the Societas Europaea Mammalogica and Tony Mitchell-Jones for providing the distributional data used to prepare the *Atlas of European mammals*. We are grateful to Marc Metzger for resampling the EnS data to the mammal grid, to Raino Lampinen for help with transforming the mammal data to the coordinate grid system and to Kari Lintulaakso for data on mammal diets. Furthermore, we thank Jaakko Hollmén for the implementation of the probabilistic EM algorithm. We also thank Jukka Jernvall, Ilkka Hanski, Risto Väisänen, Juha Merilä and Alistair Evans for discussion and constructive comments.

REFERENCES

Alroy, J., Marshall, C.R., Bambach, R.K., Bezusko, K., Foote, M., Fürsich, F.T., Hansen, T.A., Holland, S.M., Ivany, L.C., Jablonski, D., Jacobs, D.K., Jones, D.C., Kosnik, M.A., Lidgaard, S., Low, S., Miller, A.I., Novacek-Gottshall, P.M., Olszewski, T.D., Ptazkowsky, M.E., Raup, D.M., Roy, K., Sepkoski, J.J., Jr, Sommers, M.G., Wagner, P.J. & Webber, A. (2001) Effects of sampling standardization on estimates of Phanerozoic marine diversification. *Proceedings of the National Academy of Sciences USA*, **98,** 6261–6266.

Badgley, C., Bartels, W., Morgan, M., Behrensmeyer, A. & Raza, S. (1995) Taphonomy of vertebrate assemblages from the Paleogene of northwestern Wyoming and the Neogene of northern Pakistan. *Palaeogeography, Palaeoclimatology, Palaeoecology*, **115,** 157–180.

Barnosky, A.D. & Carrasco, M.A. (2002) Effects of Oligo-Miocene global climate changes on mammalian species richness in the northwestern quarter of the USA. *Evolutionary Ecology Research*, **4,** 811–841.

Barnosky, A.D., Hadly, E.A. & Bell, C.J. (2003) Mammalian response to global warming on varied temporal scales. *Journal of Mammalogy*, **84,** 354–368.

Bernor, R.L. (1983) Geochronology and zoogeographic relationships of Miocene Hominoidea. *New interpretations of ape and human ancestry* (ed. by R.L. Ciochon and R.S. Corruccini), pp. 21–64. Plenum Press, New York.

Bernor, R.L. (1984) A zoogeographic theater and a biochronologic play: the time/biofacies phenomena of Eurasian and African Miocene mammal provinces. *Paléobiologie Continentale*, **14,** 121–142.

Bernor, R.L., Andrews, P.J., Solounias, N. & Van Couvering, J.A.H. (1979) The evolution of 'Pontian' mammal faunas: some zoogeographic, palaeoecologic and chronostratigraphic considerations. *Annales Géologiques Pays Hellèniques, Tome hors série*, **1,** 81–89.

Bernor, R.L., Fahlbusch, V., Andrews, P., Bruijn, H., de Fortelius, M., Rögl, F., Steininger, F.F. & Werdelin, L. (1996) The evolution of Western Eurasian Neogene mammal faunas: a chronologic, systematic, biogeographic and paleoenvironmental synthesis. *The evolution of western Eurasian Neogene mammal faunas* (ed. by R.L. Bernor, V Fahlbusch and H.W. Mittmann), pp. 449–471. Columbia University Press, New York.

Bonan, G.B., Levis, S., Kergoat, L. & Oleson, K.W. (2002) Landscapes as patches of plant functional types: an integrating concept for climate and ecosystem models. *Global Biogeochemical Cycles*, **16,** 1–11.

Cadez, I.V., Gaffney, S. & Smyth, P. (2000) A general probabilistic framework for clustering individuals and objects. *Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (ed. by R. Ramakrishnan and S. Stolfo), pp. 140–149. ACM Press, New York.

Chase, J.M. (2005) Towards a really unified theory for metacommunities. *Functional Ecology*, **19,** 182–186.

Damuth, J. (1982) Analysis of the preservation of community structure in assemblages of fossil mammals. *Paleobiology*, **8,** 434–446.

Duda, R.O., Hart, P.E. & Stork, D.G. (2000) *Pattern classification*, 2nd edn. John Wiley & Sons, New York.

Everitt, B.S. & Hand, D.J. (1981) *Finite mixture distributions*. Chapman & Hall, London.

Fortelius, M., Eronen, J.T., Jernvall, J., Liu, L., Pushkina, D., Rinne, J., Tesakov, A., Vislobokova, I.A., Zhang, Z. & Zhou, L. (2002) Fossil mammals resolve regional patterns of Eurasian climate change over 20 million years. *Evolutionary Ecology Research*, **4,** 1005–1016.

Fortelius, M., Eronen, J.T., Liu, L., Pushkina, D., Tesakov, A., Vislobokova, I.A. & Zhang, Z. (2004) Continental-scale hypsodonty patterns, climatic palaeobiogeography, and dispersal of Eurasian Neogene large mammal herbivores. *Distribution and Migration of Tertiary Mammals in Eurasia. A volume in honour of Hans de Bruijn* (ed. by J.W.F. Reumer and W. Wessels), pp. 1–11. Deinsea, Utrecht.

Hagmeier, E.M. (1966) A numerical analysis of the distributional patterns of North American mammals. II. Re-evaluation of the provinces. *Systematic Zoology*, **15,** 279–299.

Hagmeier, E.M. & Stults, D. (1964) A numerical analysis of the distributional patterns of North American mammals. *Systematic Zoology*, **13,** 125–155.

Hand, D., Mannila, H. & Smyth, P. (2001) *Principles of data mining*. MIT Press, Cambridge, MA.

Hanski, I. (1999) *Metapopulation ecology*. Oxford University Press, Oxford.

Hanski, I. & Caggiotti, O.E. (2004) *Ecology, genetics, and evolution of metapopulations*. Elsevier Academic Press, Amsterdam.

Hanski, I. & Gyllenberg, M. (1997) Uniting two general patterns in the distribution of species. *Science*, **275,** 397–400.

Hewitt, G. (2000) The genetic legacy of the Quaternary ice ages. *Nature*, **405,** 907–913.

Hijmans, R.J., Cameron, S.E., Parra, J.L., Jones, P.G. & Jarvis, A. (2005). Very high resolution interpolated climate surfaces for global land areas. *International Journal of Climatology*, **25,** 1965–1978.

Hodgson, J.G., Wilson, P.J., Grime, J.P. & Thompson, K. (1999) Allocating C-S-R plant functional types: a soft approach to a hard problem. *Oikos*, **85,** 282–294.

Holt, R.D. & Keitt, T.H. (2005) Species' borders: a unifying theme in ecology. *Oikos*, **108,** 3–6.

Holyoak, M., Leibold, M.A. & Holt, R.D. (2005) *Metacommunities: spatial dynamics and ecological communities*. University of Chicago Press, Chicago.

IUCN (2004) *2004 IUCN Red List of Threatened Species*. http://www.iucnredlist.org [accessed on 8 December, 2006].

Järvinen, O. (1979) Geographical gradients of stability in European bird communities. *Oecologia*, **38,** 51–69.

Järvinen, O. & Väisänen, R.A. (1973) Species diversity of Finnish birds. I: Zoogeographical zonation based on land birds. *Ornis Fennica*, **50,** 93–119.

Järvinen, O. & Väisänen, R.A. (1980) Quantitative biogeography of Finnish land birds as compared with regionality in other taxa. *Annales Zoologici Fennici*, **17,** 67–85.

Jernvall, J. & Fortelius, M. (2002) Common mammals drive the evolutionary increase of hypsodonty in the Neogene. *Nature*, **417,** 538–540.

Jernvall, J. & Fortelius, M. (2004) Maintenance of trophic structure in fossil mammal communities: site occupancy and taxon resilience. *The American Naturalist*, **164,** 614–624.

Kleinberg, L. & Tardos, E. (2005) *Algorithm design*. Addison-Wesley, Boston, MA.

Krzanowski, W. J. (1988) *Principles of multivariate analysis*. Oxford University Press, New York.

Laurent, J.-M., Bar-Hen, A., Francois, L., Ghislain, M. & Cheddadi, R. (2004) Refining vegetation simulation models: from plant functional types to bioclimatic affinity groups of plants. *Journal of Vegetation Science*, **15,** 739–764.

Leibold, M.A. & Miller, T.E. (2004) From metapopulations to metacommunities. *Ecology, genetics, and evolution of metapopulations* (ed. by I. Hanski and O.E. Caggioitti), pp. 133–150. Elsevier/Academic Press, Amsterdam.

McGill, B.J., Hadly, E.A. & Maurer, B.A. (2005) Community inertia of Quaternary small mammal assemblages in North America. *Proceedings of the National Academy of Sciences USA*, **102,** 16701–16706.

McLachlan, G & Peel, D. (2000) *Finite mixture models*. John Wiley & Sons, New York.

Metzger, M.J., Bunce, R.G.H., Jongman, R.H.G., Mücher, C.A. & Watkins, J.W. (2005) A climatic stratification of the environment of Europe. *Global Ecology and Biogeography*, **14,** 549–563.

Mitchell-Jones, A.J., Amori, G., Bogdanowicz, W., Krystufek, B., Reijnders, P.J.H., Spitzenberger, F., Stubbe, M., Thissen, J.B.M., Vohralik, V. & Zima, J. (1999) *The atlas of European mammals*. Academic Press, London.

Monserud, R.A. & Leemans, R. (1992) The comparison of global vegetation maps. *Ecological Modelling*, **62,** 275–296.

Nowak, R.M. (1999) *Walker's mammals of the world*, 6th edn. The Johns Hopkins University Press, Baltimore, MD.

Olson, E.C. (1951) The evolution of a Permian vertebrate chronofauna. *Evolution*, **6,** 181–196.

Rencher, A.C. (2002) *Methods of multivariate analysis*, 2nd edn. John Wiley & Sons, New York.

Theodoridis, S. & Koutroumbas, K. (2003) *Pattern recognition*, 2nd edn. Elsevier Academic Press, New York.

Udvardy, M.D.F. (1969) *Dynamic zoogeography*. Van Nostrand Reinhold, London.

Vermeij, G.J. & Herbert, G.S. (2004) Measuring relative abundance in fossil and living assemblages. *Paleobiology*, **30,** 1–4

Wallace, A.R. (1876) *The geographical distribution of animals*. Harper, New York.

SUPPLEMENTARY MATERIAL

The following supplementary material is available for this article:

Table S1 The presence/absence records of mammals in the 50 km UTM gird and the values of the environmental variables

Table S2 The conservation status and diet data for the mammal species used in this study

Table S3 Results and *P*-values of the ANOVA tests

This material is available as part of the online article from: http://www.blackwell-synergy.com/http://www.blackwell-synergy.com/doi/abs/10.1111/j.1365-2699.2006.01664.x (this link will take you to the article abstract).

Please note: Blackwell Publishing are not responsible for the content or functionality of any supplementary materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.

BIOSKETCHES

Hannes Heikinheimo is working on his doctoral thesis under the supervision of Heikki Mannila. His interests include combinatorial and probabilistic methods for data mining with a focus on issues related to spatial and ordinal relationships in data.

Mikael Fortelius is a palaeontologist specializing in the fossil mammal faunas of Eurasia during the last 20 million years. His main interests include the relationship between environment and community structure at evolutionary time-scales and the dietary evolution of hoofed mammals.

Jussi Eronen is a palaeontologist with a wide range of interests, from biogeography and climate to palaeoecology of communities. His research focuses on the evolution and dynamics of large mammal herbivore communities in the deep time of Eurasia.

Heikki Mannila is a computer scientist working in algorithms for data analysis. His main interests are in data mining methods for 0–1 data, especially pattern discovery, sequence segmentation and spatiotemporal data analysis.

Editor: Bradford Hawkins