Data analysis of small, large, small-and-large mammals
Ella Bingham, 28th October 2005
data_analysis.tex

Results on the data set smallmammals030717_w20_lasm_spc.txt that contains 6969 rows (excluding the title), of which 3861 are "small" and 3108 are "large". For comparison, the data set NOW_public_030717_init_nz_lala_euro_space.txt, used in the manuscript "Spectral ordering and biochronology of European fossil mammals" contains 5299 rows, all of which are "large".

The data is pruned such that some outlier (in terms of their spectral coefficient) sites are removed; the same data sets are used for spectral analysis and thus spectral-related preprocessing has been done. Spectral-related data analysis will be described in another document: spectral_results.tex

One of the general conclusions is that parameter values 2-2 (meaning that first, taxa having at least 2 occurrences are chosen, and from that subset of data, sites having at least 2 occurrences are chosen) are the most interesting, as other data sets are not very coherent in some respects.

1 Genus-level data

1.1 Lifespans

In Table 1, the lifespans of taxa are computed as the number of MN units between (and including) the first and last appearance of the taxon. (Here MN unit 8 is *not* taken into account, such that if a taxon occurs in MN units 7 and 9, we conclude that its lifespan is 2 units.) We can conclude that small mammals clearly have a longer lifespan.

Table 1 also shows the average MN classes in the data sets. It is seen that the data sets are quite coherent in this respect.

group	gl	sl	gn	sn	μ lifespan	σ^2 lifespan	μ MNclass	σ^2 MNclass
small	10	10	99	126	7.0909	13.757	9.0238	18.5034
large	10	10	102	69	4.5686	4.1487	8.7536	19.1884
small	5	5	131	196	6.5191	13.8977	9	20.2872
large	5	5	154	162	4.6818	5.2249	9.1667	19.941
small	2	2	200	327	5.1900	14.5567	8.1529	22.3937
large	2	2	228	334	4.0526	6.2351	8.2478	18.5582

Table 1: gl = genuslimit, sl = generaatsite, gn = number of genera, sn = number of sites. Means and variances of lifespans of taxa in MN units. Means and variances of MNclasses of sites.

Figure 1 shows the distributions of the lifespans of genera, measured in MN units. The "both" data sets are not shown here, as the number of genera in "both" is about double as big as in "small" or "large", and the figures would become difficult to compare.

For small genera, there are 2 groups, one having a lifespan of approximately 5 MN units, and another 13 MN units. Large genera do not group in such a way. At 2-2, the grouping of small genera is not seen anymore, as rarer genera enter into the data. Unfortunately it is hard to give a numerical assessment whether the data is truly bimodal: at 10-10, 5-5 and 2-2, the Davies-Bouldin index for clustering the lifespans suggests many more than 2 clusters for both small and large genera. Anyway, the groups of small genera in 10-10 are, excluding the genera falling in between the groups (having lifespan of exactly 9 or 10 MN units):

- *lifespan less than 9 MN units:* 'Anomalomys' 'Eumyarion' 'Eurolagus' 'Lagopsis' 'Megacricetodon' 'Myoglis' 'Proscapanus' 'Albanensia' 'Amphechinus' 'Bransatoglis' 'Dinosorex' 'Glirudinus' 'Lanthanotherium' 'Paraglirulus' 'Heterosorex' 'Plesiodimylus' 'Cricetodon' 'Neocometes' 'Anchitheriomys' 'Ligerimys' 'Prodryomys' 'Cricetulodon' 'Crusafontina' 'Hispanomys' 'Kowalskia' 'Postpalerinaceus' 'Armantomys' 'Progonomys' 'Dipoides' 'Eliomys' 'Occitanomys' 'Huerzelerimys' 'Hystrix' 'Parapodemus' 'Alilepus' 'Archaeodesmana' 'Blarinella' 'Ruscinomys' 'Petenyia' 'Apocricetus' 'Rhagapodemus' 'Stephanomys' 'Apodemus' 'Blan-

comys' 'Rotundomys' 'Melissiodon' 'Peridyromys' 'Pseudodryomys' 'Eozapus' 'Pliopetaurista' 'Prospalax' 'Episoriculus' 'Mimomys' 'Trischizolagus' 'Castillomys' 'Chainodus' 'Plesiosorex' 'Palaeosciurus' 'Hypolagus' 'Trilophomys' 'Blarinoides' 'Beremendia' 'Galemys' 'Eucricetodon' 'Paratalpa' 'Plesiosminthus' 'Pseudotheridomys' 'Rhodanomys' 'Ritteneria'

- *lifespan more than 10 MN units:* 'Chalicomys' 'Democricetodon' 'Galerix' 'Microdyromys' 'Miodyromys' 'Prolagus' 'Glirulus' 'Miosorex' 'Muscardinus' 'Paenelimnoecus' 'Talpa' 'Trogontherium' 'Miopetaurista' 'Heteroxerus' 'Sciurus' 'Sorex' 'Pseudocricetus' 'Desmanella' 'Atlantoxerus' 'Glis' 'Paraethomys' 'Vasseuromys' 'Hylopetes' 'Blackia'

and the groups of small genera in 5-5 are, almost similarly:

- *lifespan less than 9 MN units:* 'Anomalomys' 'Eumyarion' 'Eurolagus' 'Lagopsis' 'Megacricetodon' 'Mioechinus' 'Myoglis' 'Proscapanus' 'Albanensia' 'Amphechinus' 'Dinosorex' 'Glirudinus' 'Lanthanotherium' 'Lartetium' 'Mygalea' 'Paraglirulus' 'Heterosorex' 'Plesiodimylus' 'Cricetodon' 'Florinia' 'Neocometes' 'Soricella' 'Anchitheriomys' 'Ligerimys' 'Metacordylodon' 'Prodryomys' 'Palaeotragus' 'Cricetulodon' 'Crusafontina' 'Hispanomys' 'Kowalskia' 'Postpalerinaceus' 'Myomimus' 'Progonomys' 'Dipoides' 'Eliomys' 'Occitanomys' 'Hystrix' 'Alilepus' 'Archaeodesmana' 'Blarinella' 'Ruscinomys' 'Petenyia' 'Apocricetus' 'Rhagapodemus' 'Stephanomys' 'Apodemus' 'Blancomys' 'Amblycoptus' 'Rotundomys' 'Karydomys' 'Melissiodon' 'Pentabuneomys' 'Peridyromys' 'Pseudodryomys' 'Eozapus' 'Epimeriones' 'Graphiurops' 'Pliopetaurista' 'Prospalax' 'Schizogalerix' 'Episoriculus' 'Mimomys' 'Trischizolagus' 'Castillomys' 'Chainodus' 'Plesiosorex' 'Deinsdorfia' 'Microtodon' 'Tamias' 'Palaeosciurus' 'Hypolagus' 'Crocidosorex' 'Erinaceus' 'Trilophomys' 'Blarinoides' 'Desmanodon' 'Beremendia' 'Oryctolagus' 'Micromys' 'Protatera' 'Myxomygale' 'Amphilagus' 'Forsythia' 'Galemys' 'Eucricetodon' 'Paratalpa' 'Piezodus' 'Plesiosminthus' 'Pseudotheridomys' 'Rhodanomys' 'Titanomys' 'Dimylus' 'Geotrypus' 'Ritteneria' 'Ratufa'

- *lifespan more than 10 MN units:* 'Chalicomys' 'Democricetodon' 'Galerix' 'Microdyromys' 'Miodyromys' 'Prolagus' 'Glirulus' 'Miosorex' 'Muscardinus' 'Paenelimnoecus' 'Talpa' 'Trogontherium' 'Scaptonyx' 'Miopetaurista' 'Heteroxerus' 'Sciurus' 'Sorex' 'Armantomys' 'Pseudocricetus' 'Desmanella' 'Atlantoxerus' 'Parapodemus' 'Glis' 'Paraethomys' 'Vasseuromys' 'Hylopetes' 'Blackia'
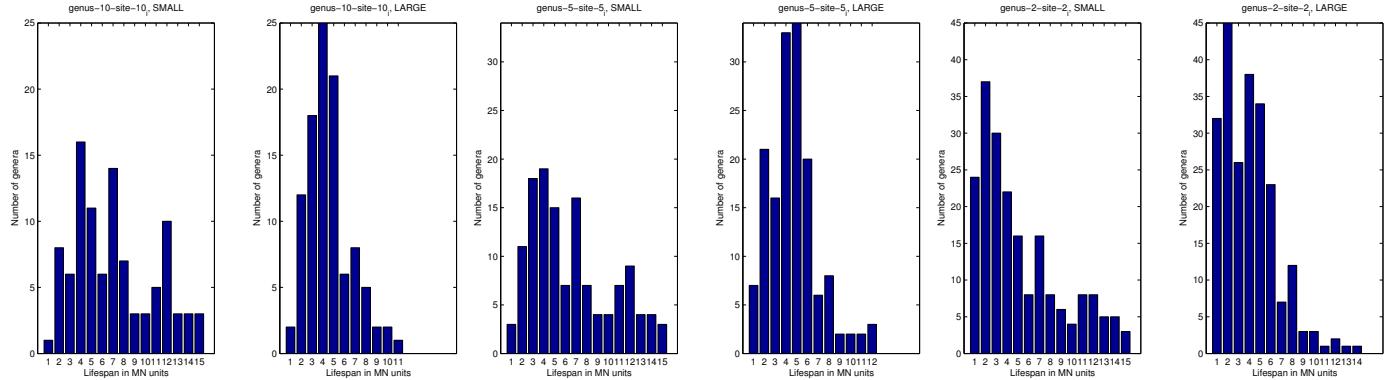

Figure 1: Distributions of lifespans of genera (in MN units): 10-10, 5-5, 2-2

We also measure the lifespans of genera in millions of years. Figure 2 shows the distributions of lifespans. For small genera, there are clearly 2 groups, and we are now able to justify this using the Davies-Bouldin index: between cluster numbers 2 and 8, the optimal choice is 2 clusters in the 10-10, 5-5 and 2-2 data sets of small genera. For large genera, the optimal choice is may more clusters in these data sets. The groups of small genera in 10-10 are, excluding the genera falling between the groups (having lifespan 8 to 11 million years):

- *lifespan less than 8 million years:* 'Eurolagus' 'Lagopsis' 'Albanensia' 'Lanthanotherium' 'Cricetodon' 'Neocometes' 'Anchitheriomys' 'Ligerimys' 'Prodryomys' 'Cricetulodon' 'Crusafontina' 'Hispanomys' 'Kowalskia' 'Postpalerinaceus' 'Armantomys' 'Progonomys' 'Dipoides' 'Eliomys' 'Occitanomys' 'Huerzelerimys'

2

'Hystrix' 'Parapodemus' 'Alilepus' 'Archaeodesmana' 'Blarinella' 'Ruscinomys' 'Petenyia' 'Apocricetus' 'Rhagapodemus' 'Stephanomys' 'Apodemus' 'Blancomys' 'Rotundomys' 'Melissiodon' 'Peridyromys' 'Eozapus' 'Pliopetaurista' 'Prospalax' 'Episoriculus' 'Mimomys' 'Trischizolagus' 'Castillomys' 'Hypolagus' 'Trilophomys' 'Blarinoides' 'Beremendia' 'Galemys' 'Eucricetodon' 'Paratalpa' 'Plesiosminthus' 'Pseudotheridomys' 'Rhodanomys' 'Ritteneria' *lifespan more than 11 million years:* 'Chalicomys' 'Democricetodon' 'Galerix' 'Microdyromys' 'Miodyromys' 'Prolagus' 'Spermophilinus' 'Bransatoglis' 'Glirudinus' 'Glirulus' 'Miosorex' 'Muscardinus' 'Paenelimnoecus' 'Steneofiber' 'Talpa' 'Trogontherium' 'Eomyops' 'Heterosorex' 'Miopetaurista' 'Heteroxerus' 'Keramidomys' 'Sciurus' 'Sorex' 'Pseudocricetus' 'Desmanella' 'Atlantoxerus' 'Glis' 'Paraethomys' 'Vasseuromys' 'Hylopetes' 'Blackia' 'Chainodus' 'Plesiosorex' 'Palaeosciurus'

and the groups of small genera in 5-5 are almost similarly:

- *lifespan less than 8 million years:* 'Eurolagus' 'Mioechinus' 'Albanensia' 'Lanthanotherium' 'Lartetium' 'Mygalea' 'Cricetodon' 'Florinia' 'Neocometes' 'Soricella' 'Anchitheriomys' 'Ligerimys' 'Metacordylodon' 'Prodryomys' 'Palaeotragus' 'Cricetulodon' 'Crusafontina' 'Hispanomys' 'Kowalskia' 'Postpalerinaceus' 'Myomimus' 'Progonomys' 'Dipoides' 'Eliomys' 'Occitanomys' 'Hystrix' 'Alilepus' 'Archaeodesmana' 'Blarinella' 'Ruscinomys' 'Petenyia' 'Apocricetus' 'Rhagapodemus' 'Stephanomys' 'Apodemus' 'Blancomys' 'Amblycoptus' 'Rotundomys' 'Karydomys' 'Melissiodon' 'Pentabuneomys' 'Peridyromys' 'Eozapus' 'Epimeriones' 'Graphiurops' 'Pliopetaurista' 'Prospalax' 'Schizogalerix' 'Episoriculus' 'Mimomys' 'Trischizolagus' 'Castillomys' 'Deinsdorfia' 'Microtodon' 'Tamias' 'Hypolagus' 'Crocidosorex' 'Erinaceus' 'Trilophomys' 'Blarinoides' 'Desmanodon' 'Beremendia' 'Oryctolagus' 'Micromys' 'Protatera' 'Myxomygale' 'Forsythia' 'Galemys' 'Eucricetodon' 'Paratalpa' 'Piezodus' 'Plesiosminthus' 'Pseudotheridomys' 'Rhodanomys' 'Titanomys' 'Dimylus' 'Geotrypus' 'Ritteneria' 'Ratufa'

- *lifespan more than 11 million years:* 'Chalicomys' 'Democricetodon' 'Galerix' 'Microdyromys' 'Miodyromys' 'Prolagus' 'Spermophilinus' 'Amphechinus' 'Bransatoglis' 'Glirudinus' 'Glirulus' 'Miosorex' 'Muscardinus' 'Paenelimnoecus' 'Steneofiber' 'Talpa' 'Trogontherium' 'Eomyops' 'Heterosorex' 'Plesiodimylus' 'Scaptonyx' 'Miopetaurista' 'Heteroxerus' 'Keramidomys' 'Sciurus' 'Sorex' 'Armantomys' 'Pseudocricetus' 'Desmanella' 'Atlantoxerus' 'Huerzelerimys' 'Parapodemus' 'Glis' 'Paraethomys' 'Vasseuromys' 'Hylopetes' 'Blackia' 'Chainodus' 'Plesiosorex' 'Palaeosciurus' 'Amphilagus'
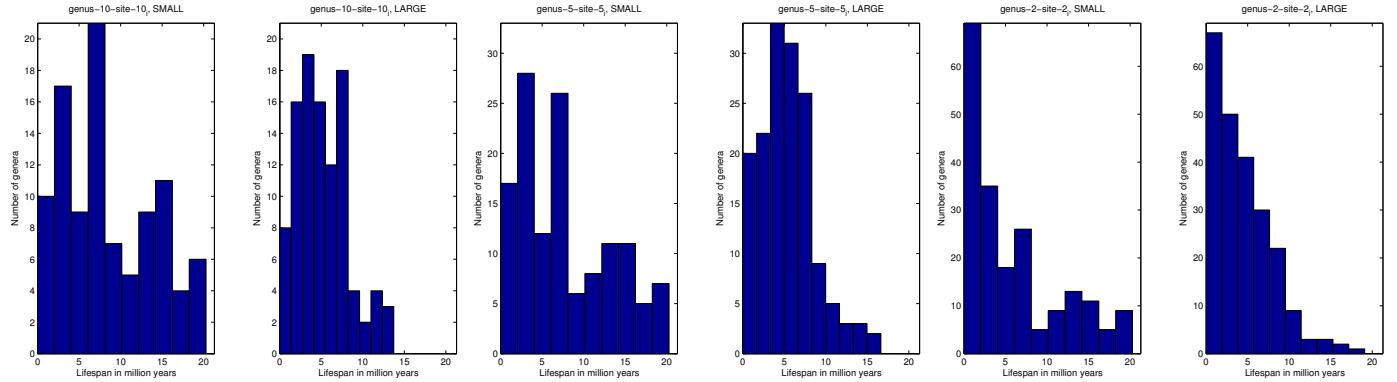

Figure 2: Distributions of lifespans of genera (in millions of years): 10-10, 5-5, 2-2

1.2 Lifespan decays

As a further analysis we show how the lifespan decays: we compute the number of genera having a lifespan of at least 1 MN unit; at least 2 MN units; at least 3 MN units; and so on. The decay rates of small and large genera are quite different, as seen in Figure 3: the lifespans of large mammals drop down more steeply.

A similar analysis is done when the lifespans are measured as millions of years. Figure 4 shows the decay of the lifespan, and figure 5 shows the same in logarithmic units, to enable comparisons with the "Van Valen" results. Furthermore, figure 6 shows the decay at short-lived small genera, long-lived small genera and large genera separately.
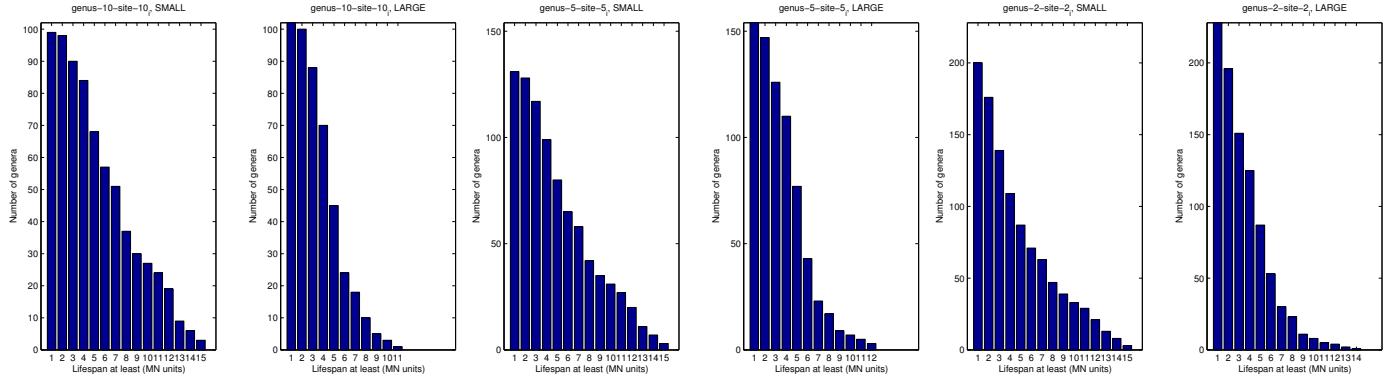
Figure 3: Lifespan decays of genera (in MN units): 10-10, 5-5, 2-2. The figures give the number of genera having a lifespan of at least 1 MN unit; at least 2 MN units; at least 3 MN units; and so on.
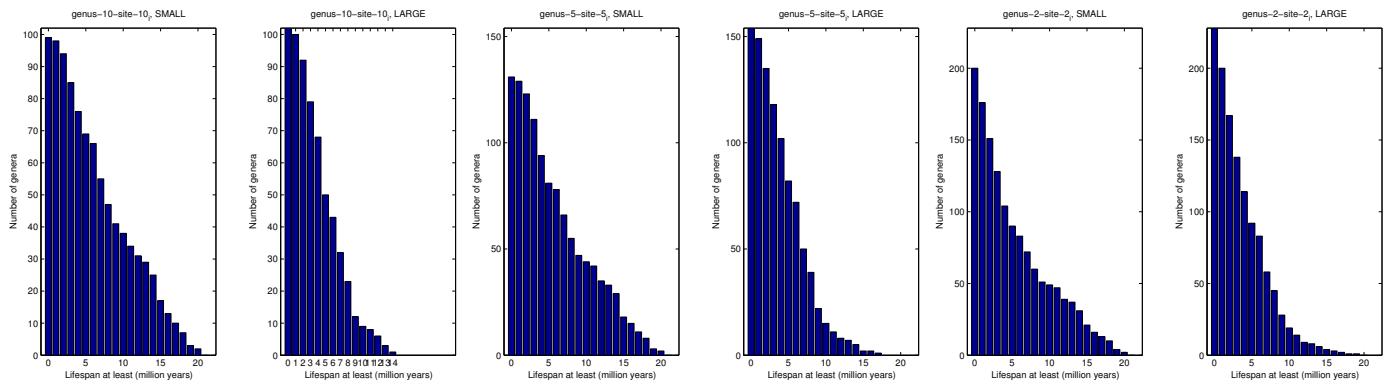

Figure 4: Lifespan decays of genera (in millions of years): 10-10, 5-5, 2-2. The figures give the number of genera having a lifespan of at least 0 million years; at least 1 million years; at least 2 million years; and so on.
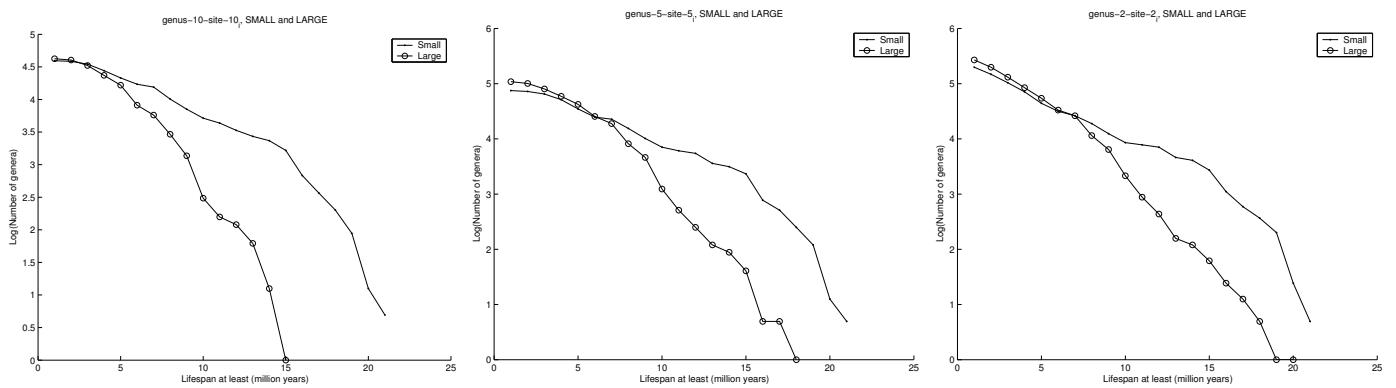

Figure 5: Lifespan decays of genera (in millions of years): 10-10, 5-5, 2-2, values of Figure 4 in logarithmic scale.
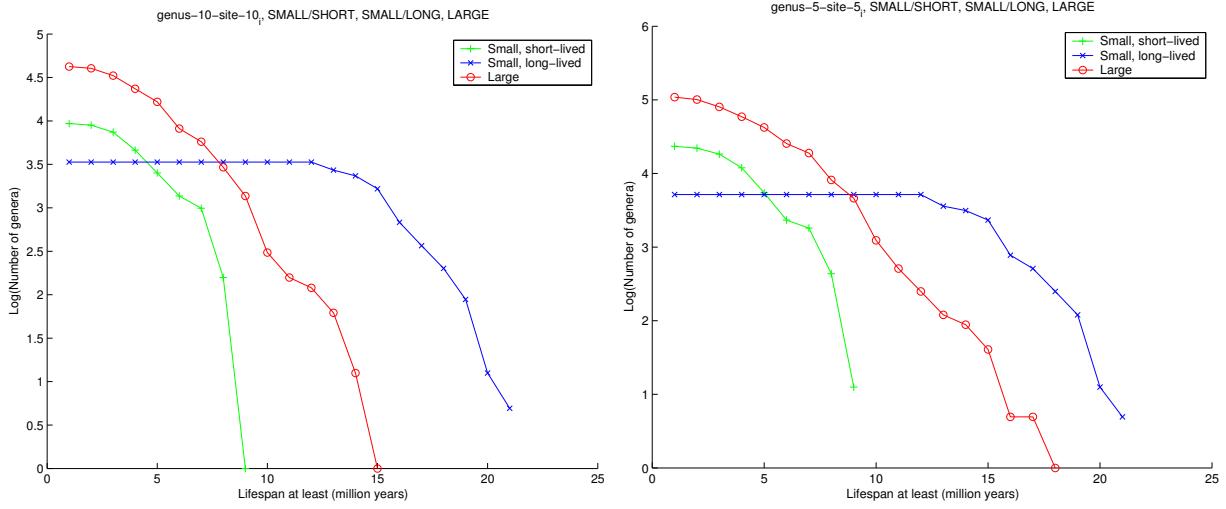
Figure 6: Lifespan decays of genera (in millions of years): 10-10, 5-5, in logarithmic scale. Small mammals are divided into short-lived (less than 8 million years; 53 genera in 10-10 and 79 in 5-5) and long-lived (more than 11 million years; 34 genera in 10-10 and 41 in 5-5), excluding intermediate (12 genera in 10-10 and 11 in 5-5). Large mammals have 102 genera in 10-10 and 154 in 5-5.

1.3 Prevalence

We then study the prevalence of genera and compare that to the lifespans. For each genus, we pick the most representative MN class as the one in which the genus has most observations. Within this MN class, we compute the prevalence as the longest distance between two sites in which the genus was observed. Furthermore, we divide the distance by the longest distance between any two sites in the MN class. The relative prevalence obtained in this way measures to which degree the genus has spread onto the whole are known at the time.

Figure 7 shows the distribution of relative prevalences of genera. Value 1 means that the genus occupies the whole area of its most representative MN class. Value 0 means that there is only one observation and no distance can be computed. The relative prevalence, computed as described above, is a very reliable measure of prevalence, as it avoids the bias of migration during a very long lifespan, and also the bias of different geographical areas of MN units. It is hard to identify any clear patterns in the figures, except that large animals have more extreme values (0 and 1). Also, at 2-2 the rare genera enter into play, making the number of 0 distances very large.
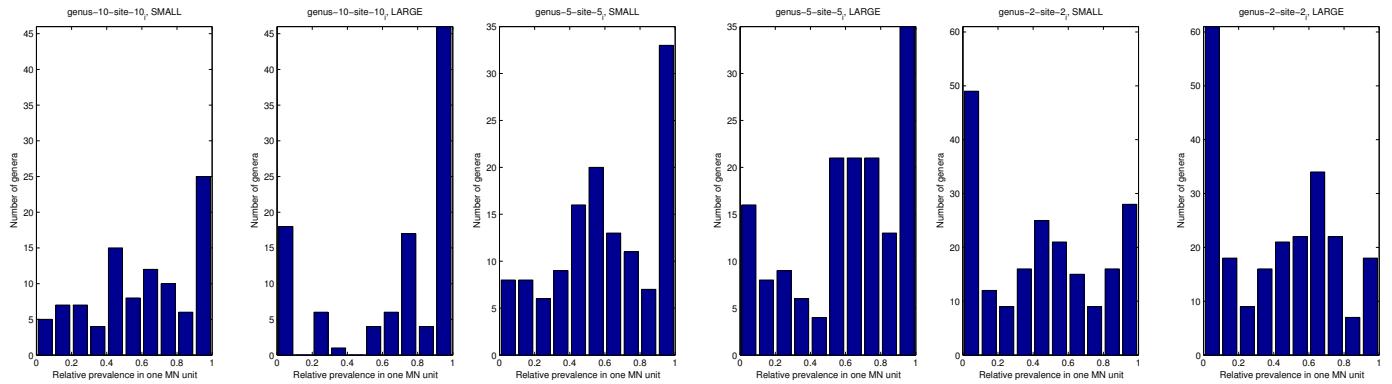
Figure 7: Distribution of the relative prevalence of genera. For each genus, the most representative MN class is first chosen; then, the distance between two observed sites is divided by the maximal distance of any two sites in that MN class.

On the other hand, ignoring the geographical areas of MN units, we get what is shown in Figure 8 where the largest distance between two observed sites is not divided by the largest distance of any two sites in the corresponding MN class. The distance is measured as the Euclidean distance of the locations expressed in

5

latitudes and longitudes. Again, there are a lot of large mammals having distance 0, meaning that they only have one occurrence in the most representative MN class. At 10-10, small mammals reach into larger distances than large mammals.
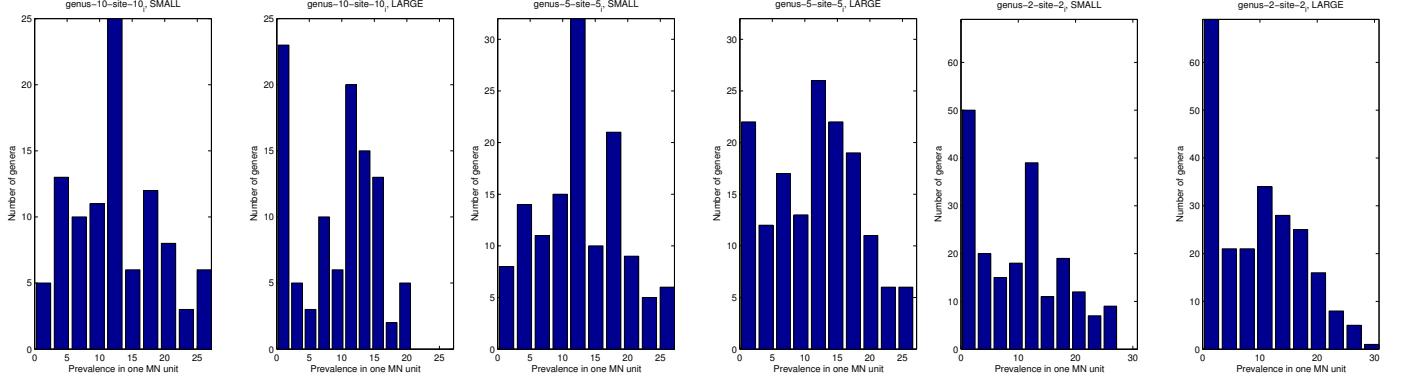

Figure 8: Distribution of the prevalence of genera. For each genus, the most representative MN class is first chosen; then, the maximum distance between two observed sites is measured.

Further, ignoring the possible bias of long-lived taxa migrating to larger areas over time, we analyze the prevalence of genera during their whole lifetime. In contrast to the above, all MN units are now taken into account. Figure 9 shows the result of such an analysis. Indeed, small mammals at 10-10 occupy the far end of the axis; at 5-5 and 2-2 this is not as clear.
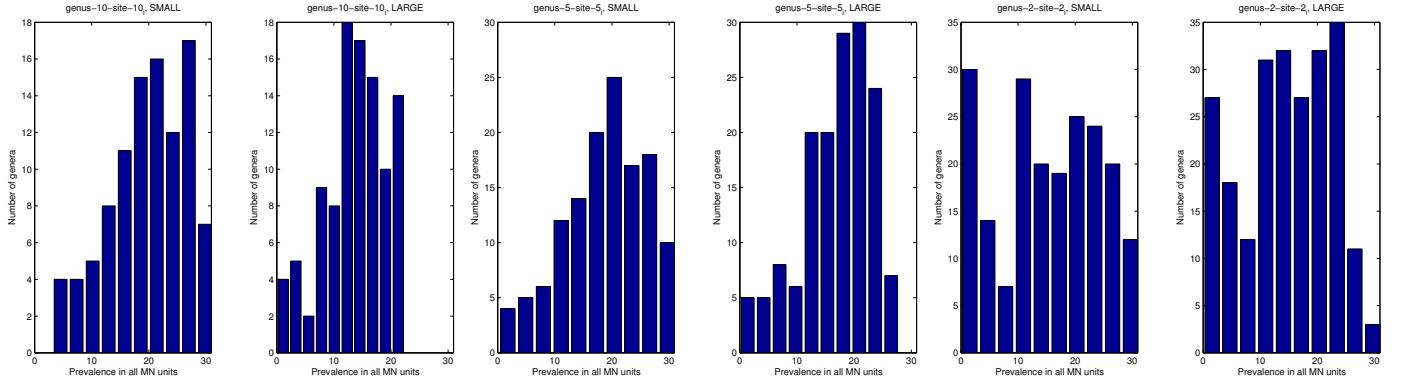
Figure 9: Distribution of the prevalence of genera. For each genus, the maximum distance between two observed sites (in any MN class) is measured.

We then measure the correlation between the prevalence and the lifespan. The correlations, listed in Table 2, show that there are basically no correlations. This is probably due to the uneven geographical locations of the sites. Also, when only one MN class is taken into account, the numbers of observations are quite small. The largest correlations are obtained at small mammals when the lifespan and prevalence in all MN units are compared — but here a possible bias exists, as the taxon might migrate over time.

6

group	gl	sl	gn	sn	corr(lifespan,rel.prev.)	corr(lifespan, prev.)	corr(lifespan,univ.prev.)
small	10	10	99	126	0.06956	0.049027	0.4931
large	10	10	102	69	0.03828	0.075246	0.30336
small	5	5	131	196	0.11317	0.1328	0.53581
large	5	5	154	162	0.03192	0.16826	0.39619
small	2	2	200	327	0.40571	0.40657	0.64798
large	2	2	228	334	0.25694	0.32763	0.55476

Table 2: gl = genuslimit, sl = generaatsite, gn = number of genera, sn = number of sites. Correlations between the lifespan and prevalence: rel.prev. = maximum distance between observations in the most representative MN class divided by the maximum distance between any two sites in the MN class, prev. = maximum distance between observations in the most representative MN class, univ.prev. = maximum distance between observations everywhere.

1.4 Community ages

Figure 10 shows the distributions of the community ages (in MN units). Here the community age is computed as follows. At a site, we study the taxa found in the site, and compute the age of a taxon as the difference between the MNclass of the site and the smallest MNclass at which the taxon is observed. The community age of the site is the average of the ages of its taxa. (Here, MNclass 8 is taken into account, so a taxon first seen at a site of MNclass 7 has age 2 at a site of MNclass 9.)

Again, the community ages of large mammals are younger than those of small mammals. Other observations include: At 2-2 we have more short-aged sites than at 5-5 and 10-10, because short-aged sites typically have only a few occurrences and they are thus excluded in 5-5 and 10-10. Also, the distribution of "both" is often different from "small" and "large" because in the "both" data set there are sites with only a few taxa, and taxa with only a few occurrences — that is, short-lived taxa.
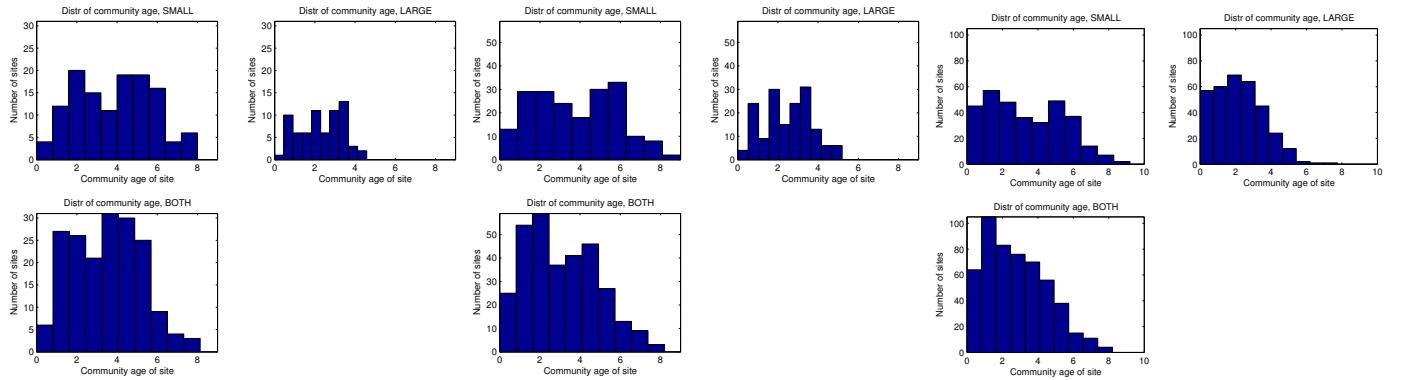
Figure 10: Distributions of community ages of sites (in MN units): 10-10, 5-5, 2-2

Similarly, Figure 11 shows the distributions of the community remaining ages (in MN units). The community remaining age is computed similarly to the community age; namely, by comparing the MNclass of the site and the largest MNclass in which a taxon is observed. Again, community remaining ages of large mammals are shorter.

The community ages and community remaining ages of MN units are shown in Figures 12 and 13. For each MN unit, the community ages (and remaining ages) of sites belonging to the MN unit are averaged. 90 per cent confidence intervals of the means of Normal distribution are also seen — if the confidence interval is not shown, there was only one age in a MN unit and no variance could be computed. This is the case at 10-10 (leftmost plot) at MN1-2 and MN14-16 "large", MN17 "small" etc. in Figure 12, and MN1-2 and MN14-17 "small" etc. in Figure 13.

In Figure 12, large mammals seem to suffer from a drastic shock before MN unit 10: at MN10 they are much younger on average than at MN9; furthermore, at MN11 they are again older. A similar (but inverse)
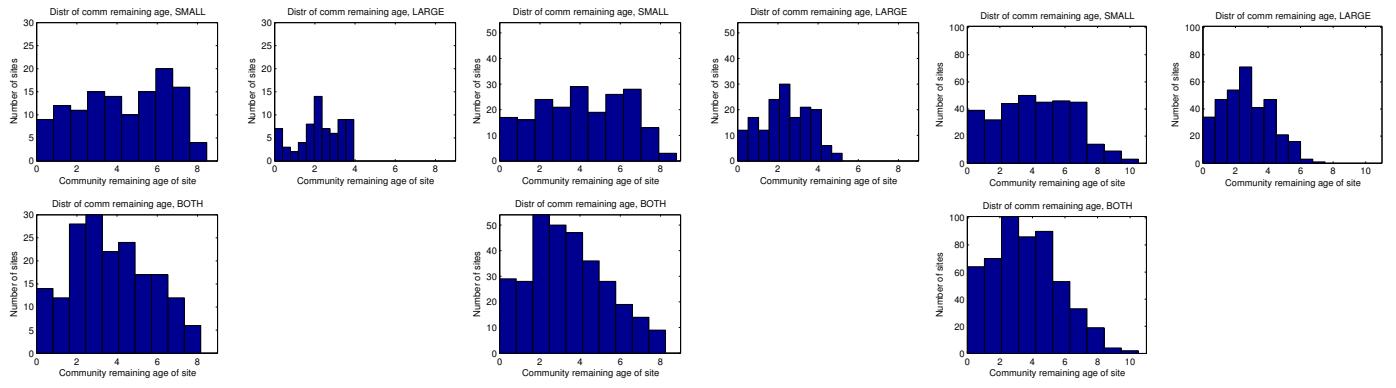
Figure 11: Distributions of community remaining ages of sites: 10-10, 5-5, 2-2

phenomenon is seen in the community remaining ages in Figure 13.

Another point perhaps worth mentioning is that the community remaining ages seem to peak at MN5 at genera (Figure 13) whereas the peak of species is at MN7 (Figure 25). Perhaps the species within a genus are distributed such that the first ones are "pathfinders" that only live for a quite short period, and the later species are more long-living. — But this is just an amateur's quick conclusion!
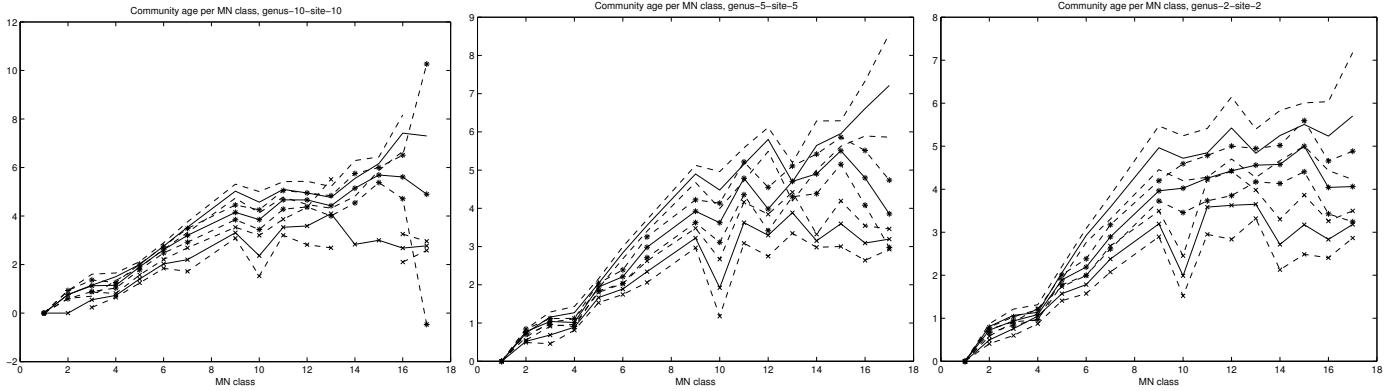
Figure 12: Community ages per MN classes: 10-10, 5-5, 2-2. '—' small, '–×–' large, '–∗–' both. Dashed lines show the 90 % confidence intervals.
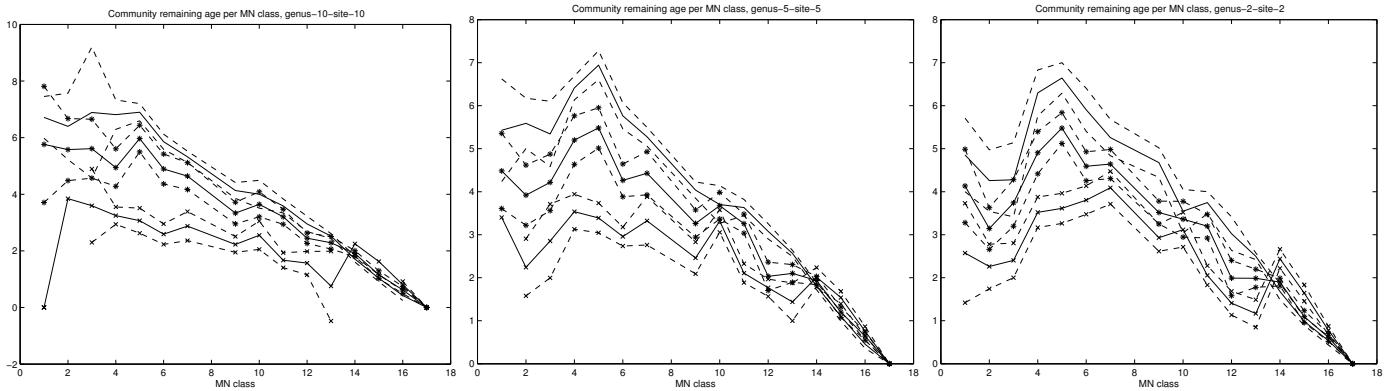
Figure 13: Community remaining ages per MN classes: 10-10, 5-5, 2-2. '—' small, '−×−' large, '−∗−' both. Dashed lines show the 90 % confidence intervals.

1.5   Other plots
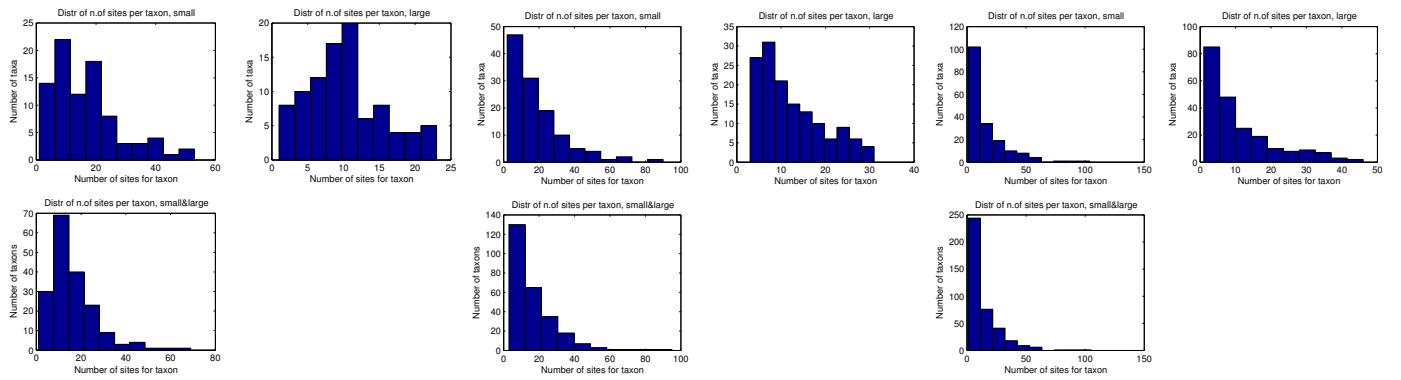
Figure 14: Distribution of number of sites per taxon. 10-10, 5-5, 2-2
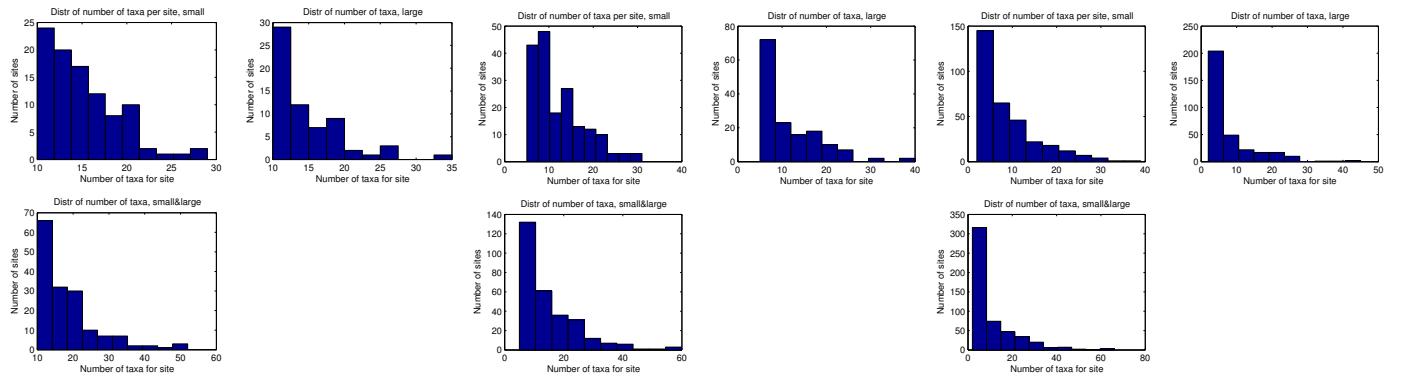
Figure 15: Distribution of number of taxa per site. 10-10, 5-5, 2-2
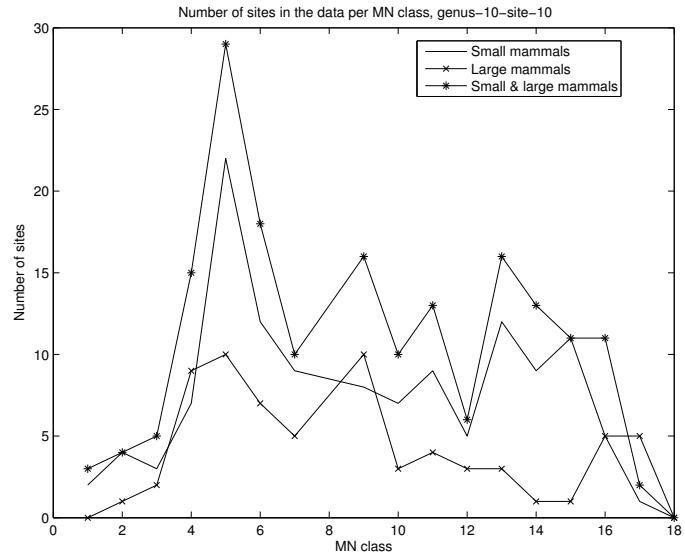
Figure 16: Number of sites per MNclass, genus-10-site-10
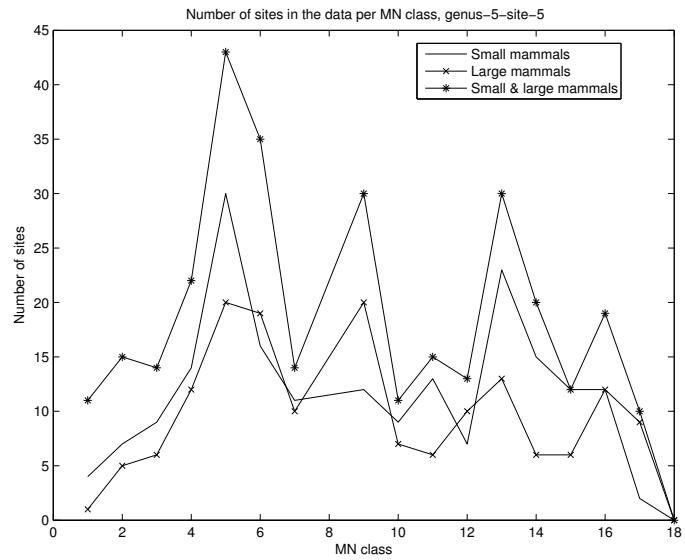
Figure 17: Number of sites per MNclass, genus-5-site-5
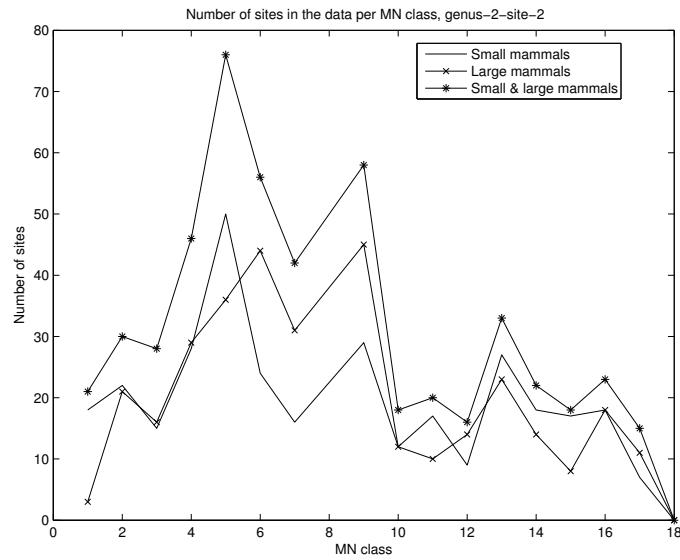
Figure 18: Number of sites per MNclass, genus-2-site-2
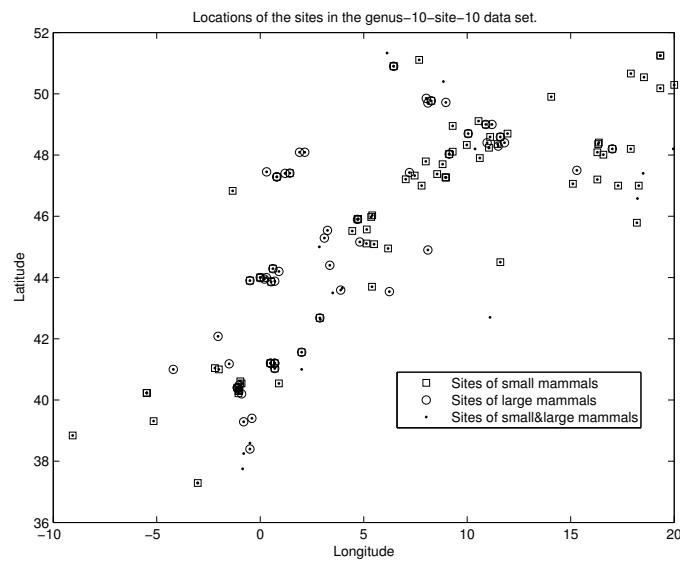
Figure 19: Locations of the sites, genus-10-site-10
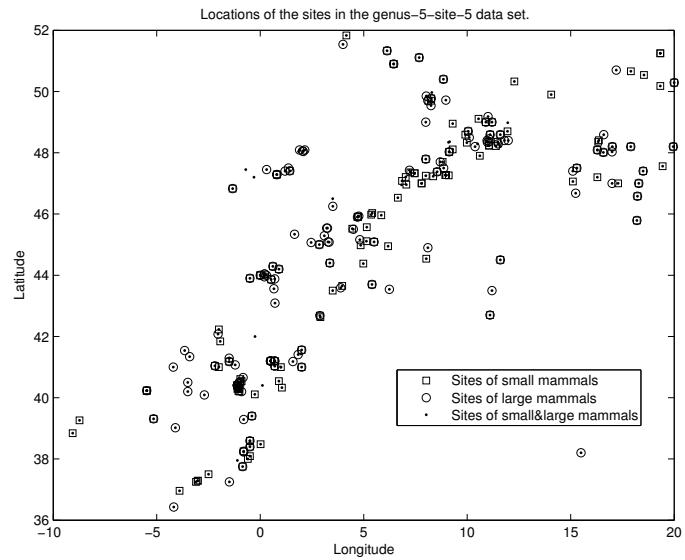
Figure 20: Locations of the sites, genus-5-site-5
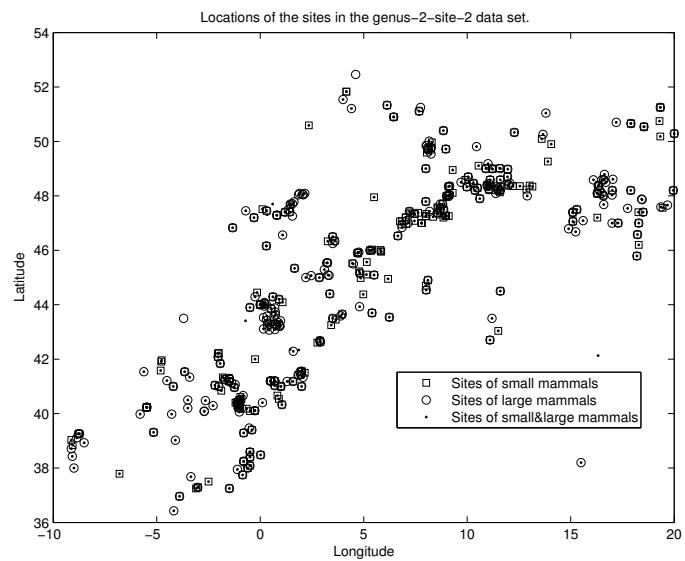
Figure 21: Locations of the sites, genus-2-site-2

2 Species-level data

2.1 Lifespans and community ages, species data

In Table 3, the lifespans of taxa are computed as the number of MN units in which they occur. Here MN unit 8 is taken into account, such that if a taxon occurs in MN units 7,8,9, we conclude that its lifespan is 3 years. Again, small species have a longer lifespan.

Table 3 also shows the average MN units of the sites. The species data sets are not as coherent in this respect as what was seen earlier in genus data in Table 1. One must note that especially the 10-10 data set is quite small and thus spurious differences are seen in the average behaviours of the data sets.

group	gl	sl	gn	sn	μ lifespan	σ^2 lifespan	μ MNclass	σ^2 MNclass
small	10	10	95	60	4.3263	7.8817	8.35	11.4178
large	10	10	61	28	3.5902	3.3459	7.2143	4.0265
both	10	10	173	101	4.2775	6.9109	8.1386	11.4406
small	5	5	239	181	4.5314	10.124	8.7569	19.8295
large	5	5	186	143	3.5	3.0838	8.8042	19.0036
both	5	5	426	282	4.1714	7.3847	8.773	20.0195
small	2	2	532	320	3.3421	7.4232	8.2	22.4928
large	2	2	418	330	2.8469	3.2811	8.1424	18.08
both	2	2	948	518	3.1487	5.7508	7.8514	19.7051

Table 3: Sizes of data sets: gl = genuslimit, sl = generaatsite, gn = number of genera, sn = number of sites. Means and variances of lifespans of taxa in MN units. Means and variances of MNclasses of sites. Species data.

Figures 22 and 23 show the distributions of community ages and remaining ages of sites. Cases 10-10 are not very informative as the corresponding data sets are quite small.
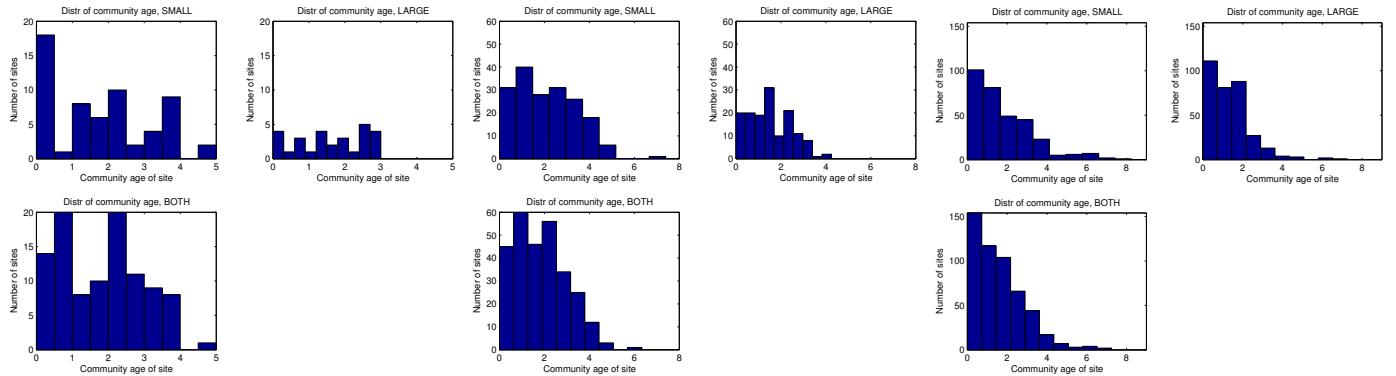

Figure 22: Distributions of community ages of sites (in MN units): 10-10, 5-5, 2-2
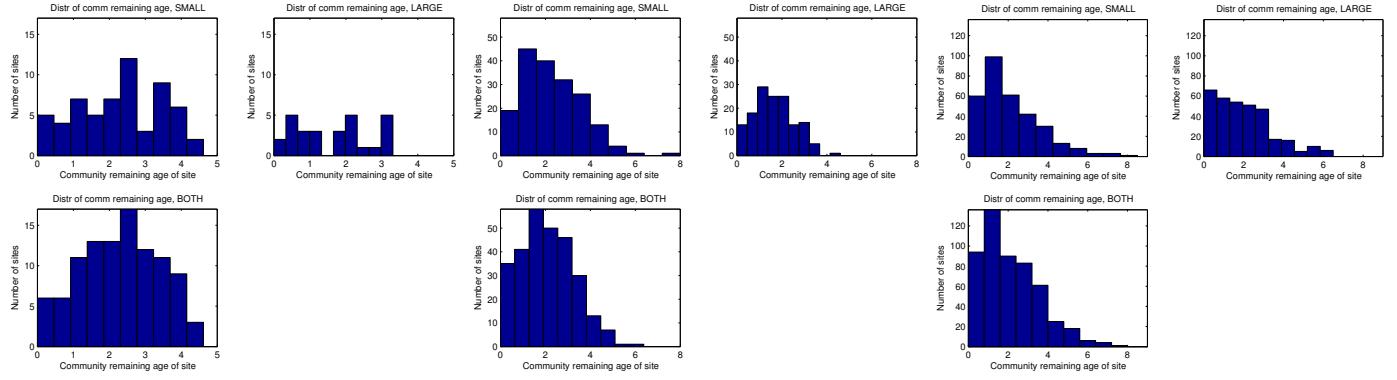
Figure 23: Distributions of community remaining ages of sites (in MN units): 10-10, 5-5, 2-2
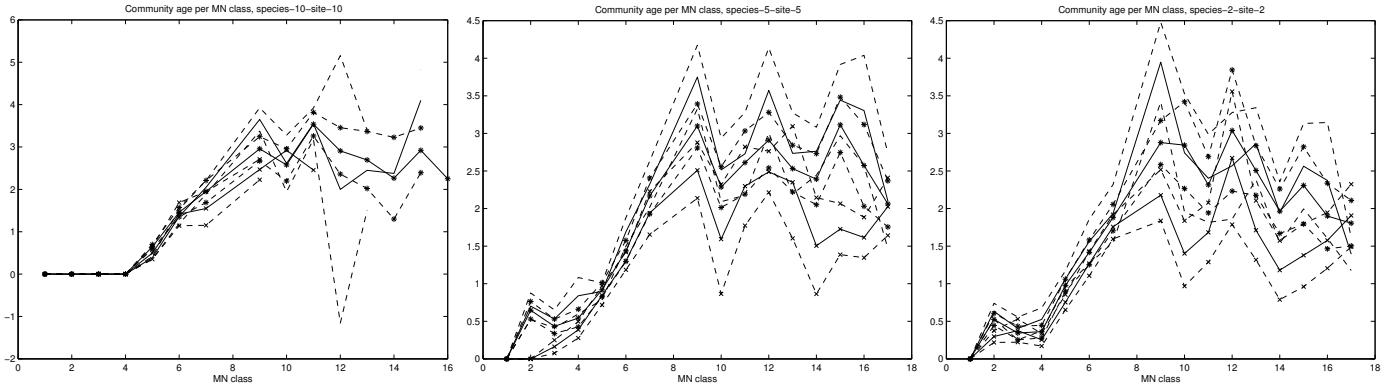

Figure 24: Community ages per MN classes: 10-10, 5-5, 2-2. '—' small, '–×–' large, '–∗–' both. Dashed lines show the 90 % confidence intervals.
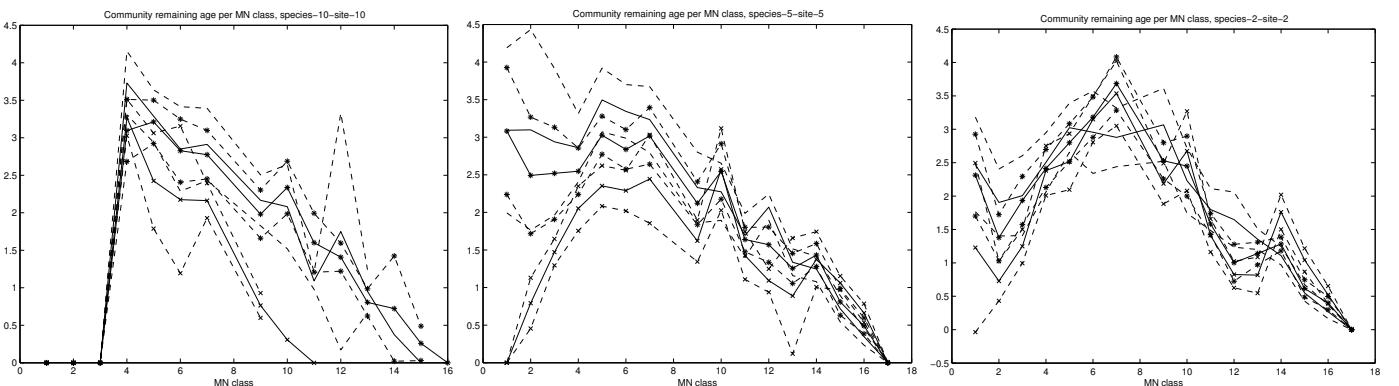

Figure 25: Community remaining ages per MN classes: 10-10, 5-5, 2-2. '—' small, '–×–' large, '–∗–' both. Dashed lines show the 90 % confidence intervals.

2.2 Other plots, species data
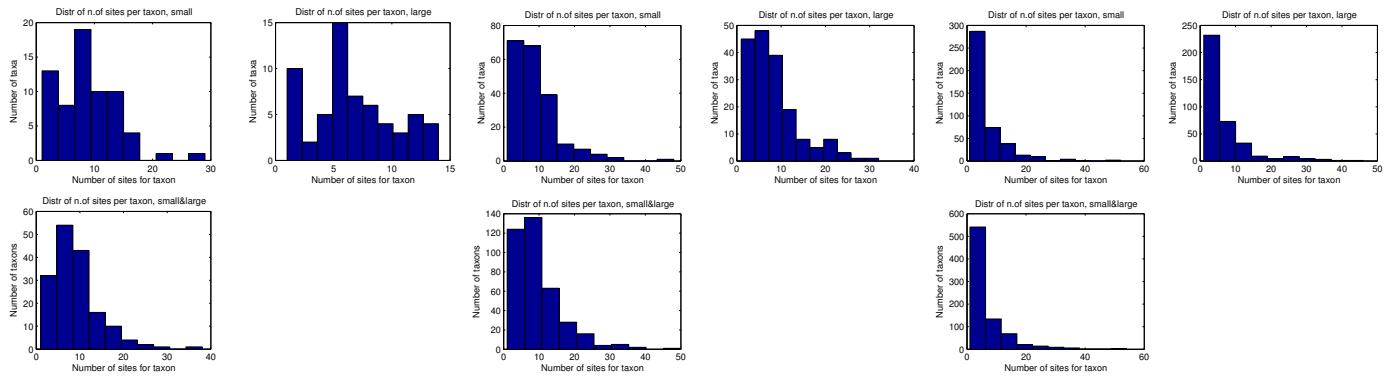
Figure 26: Distribution of number of sites per taxon. 10-10, 5-5, 2-2
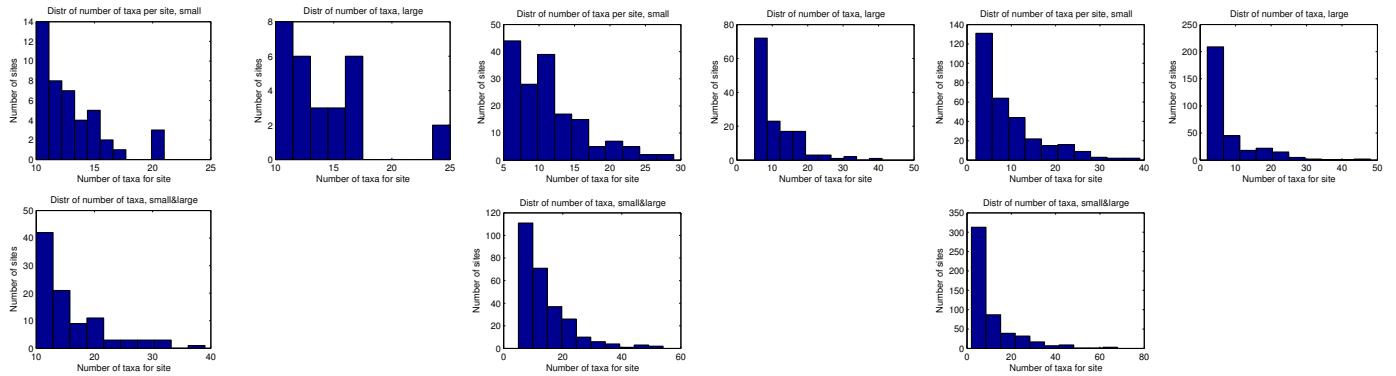
Figure 27: Distribution of number of taxa per site. 10-10, 5-5, 2-2
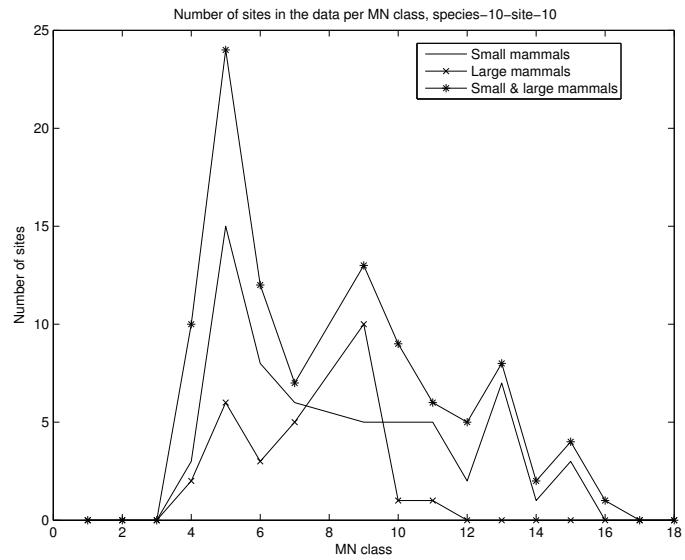
Figure 28: Number of sites per MNclass, species-10-site-10
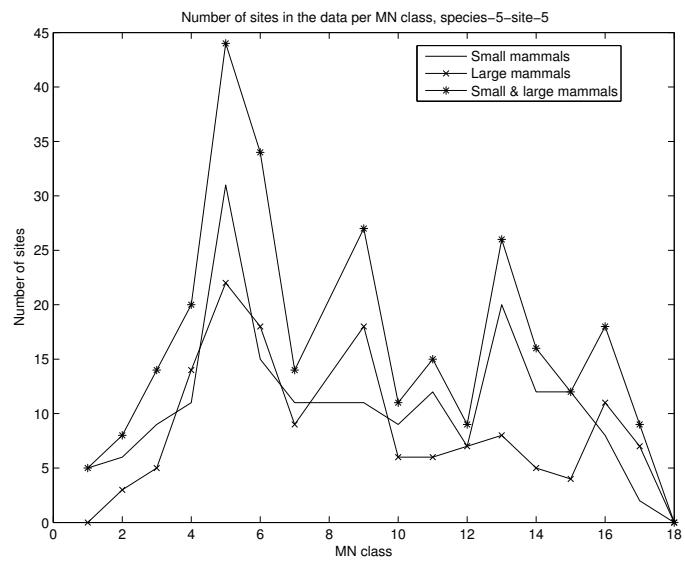
Figure 29: Number of sites per MNclass, species-5-site-5
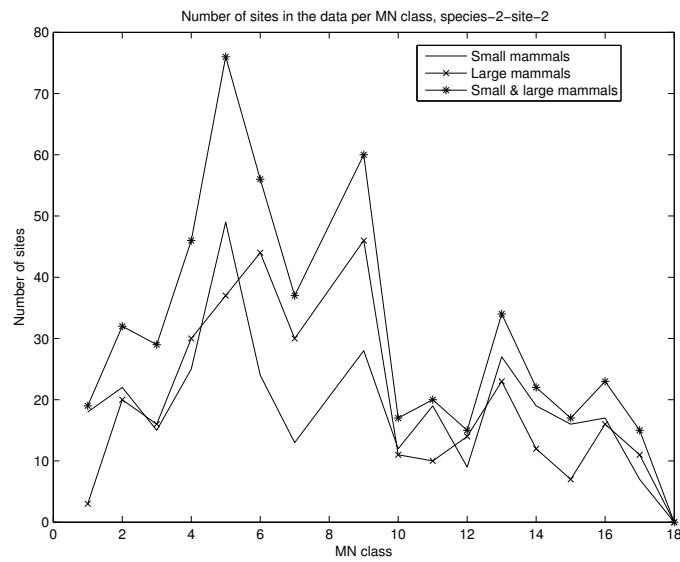
Figure 30: Number of sites per MNclass, species-2-site-2
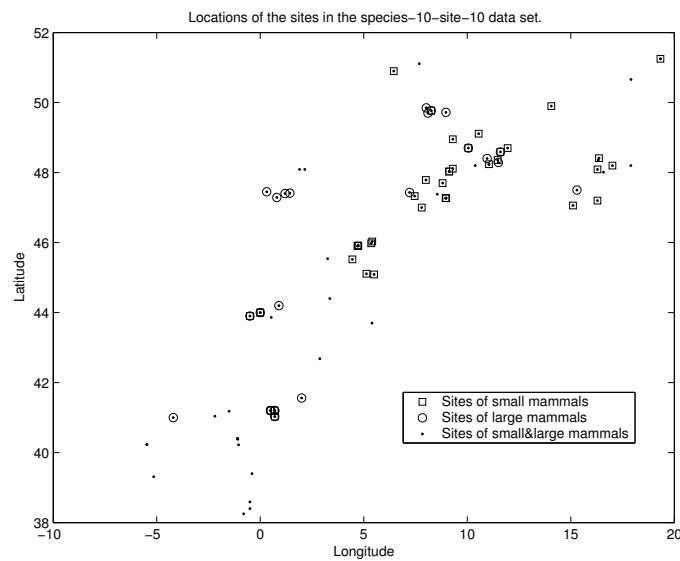
Figure 31: Locations of the sites, species-10-site-10
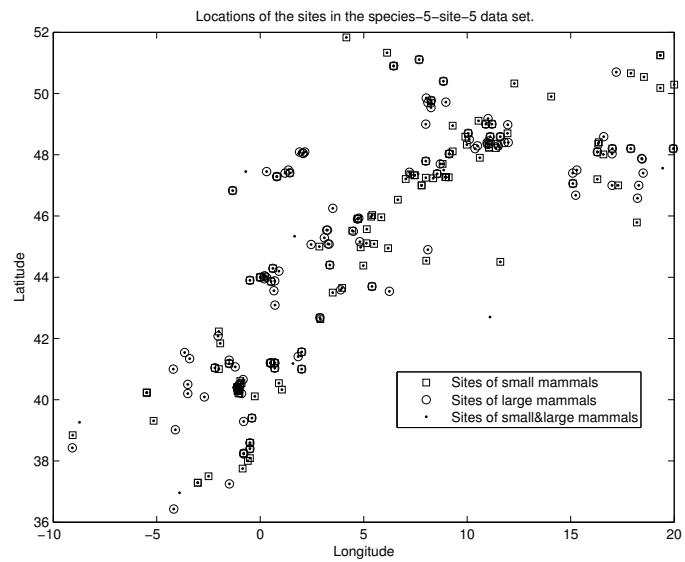
Figure 32: Locations of the sites, species-5-site-5
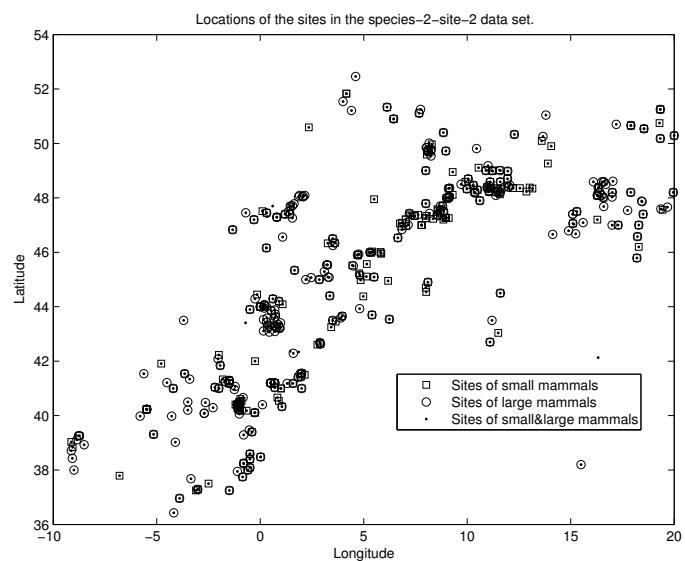
Figure 33: Locations of the sites, species-2-site-2